Towards Data Governance for International Dementia Care Mapping (DCM)

A Study Proposing DCM Data Management through a Data Warehousing Approach

Shehla KHALID

*Submitted for the degree
of Master of Philosophy*

**School of Computing,
Informatics and Media
University of Bradford**

2010

Abstract

Key Words: Dementia Care Mapping (DCM), Data Governance, Data Management, Data Warehousing

Information Technology (IT) plays a vital role in improving health care systems by enhancing the quality, efficiency, safety, security, collaboration and informing decision making. Dementia, a decline in mental ability which affects memory, concentration and perception, is a key issue in health and social care, given the current context of an aging population. The quality of dementia care is noted as an international area of concern.

Dementia Care Mapping (DCM) is a systematic observational framework for assessing and improving dementia care quality. DCM has been used as both a research and practice development tool internationally. However, despite the success of DCM and the annual generation of a huge amount of data on dementia care quality, it lacks a governance framework, based on modern IT solutions for data management, such a framework would provide the organisations using DCM a systematic way of storing, retrieving and comparing data over time, to monitor progress or trends in care quality.

Data Governance (DG) refers to the implications of policies and accountabilities to data management in an organisation. The data management procedure includes availability, usability, quality, integrity, and security of the organisation data according to their users and requirements.

This novel multidisciplinary study proposes a comprehensive solution for governing the DCM data by introducing a data management framework based on a data warehousing approach. Original contributions have been made through the design and development of a data management framework, describing the DCM international database design and DCM data warehouse architecture. These data repositories will provide the acquisition and storage solutions for DCM data. The designed DCM data warehouse facilitates various analytical applications to be applied for multidimensional analysis. Different queries are applied to demonstrate the DCM data warehouse functionality.

A case study is also presented to explain the clustering technique applied to the DCM data. The performance of the DCM data governance framework is demonstrated in this case study related to data clustering results. Results are encouraging and open up discussion for further analysis.

Acknowledgements

First of all I would like to thank God for giving me strength and courage to carry out this study.

I would like to thank my supervisor Dr. Daniel Neagu from School of Computing, Informatics and Media (SCIM) for his support, constructive criticism and encouragement throughout my MPhil study. I am grateful to him and SCIM for providing me a chance to do this MPhil study.

I would also like to thank my supervisor Dr. Claire Surr from School of Health Studies (SOH) for her continuous support, encouragement and belief in my work. She has also been very helpful in providing me the chances to participate in the DCM courses and providing me the DCM data for experimental purposes.

I would like to express my gratitude to both my supervisors to help me to write the paper for an International conference and also giving me opportunities to attend different conferences and national and international important events to present my MPhil work.

In the end I would like to thank my family for their great support, love and encouragement for me to take on new academic challenges.

Peer Reviewed Publications and Contributions

Conference Publications

S. Khalid, C. Surr and D. Neagu, (2010) "DCM Data Management Framework: A Data Warehousing Approach" in Information Technology in Bio- and Medical Informatics, ITBAM 2010, S.Khuri, L. Lhotska, N. Pisanti (Eds), LNCS: 6266 Proceedings, pp. 45- 56.

Presentations

1) "DCM Data Management Framework: A Data Warehousing Approach" at the 1st International Conference on Information Technology in Bio-and Medical Informatics, ITBAM 2010 in Bilbao, Spain on 30th August 2010.

2) Presented of the original prototype of the DCM International database system at the annual DCM International Implementation Group Meeting organised by Bradford Dementia Group on 15th November 2009.

3) Presented of progress on the DCM databases at the annual UK DCM trainers meeting, organised by Bradford Dementia Group on 29th November 2009.

4) Invited presentation on the DCM data management system to the members of Care Quality Commission (CQC), in London in (date is not decided yet) October 2010.

Training Courses

1) Attended and passed a short training course *learning to use DCM* (Basic User Status) on 23rd to 26th June 2009 organised by Bradford Dementia Group during my MPhil. This course helped me to learn about the DCM system, how it works and the information about potential users.

2) Attended and passed the DCM short course *Using DCM for practice development* (Advance DCM User Status) on 20th to 23rd October 2009 organised by Bradford Dementia Group. I learned about DCM data interpretation, its use in different DCM settings and potential end users.

Attended Conferences

1) The 21st International Conference on Database and Expert Systems Applications – DEXA4'10 in Bilbao, Spain on 29th August to 3rd September. Chaired session on 1st September 2010.

2) The international conference on eHealth in Leeds on 22nd March 2010.

3) The international conference on eHealth (Revolutionising Healthcare) in London on 12th May 2010.

Table of Contents

LIST OF FIGURES

LIST OF TABLES

Chapter 1: Introduction

Electronic health (e-health) is an emerging innovation in the health care industry. Information Technology's (IT) contribution to the health care system improves quality, efficiency, safety, security, collaboration and informs and assists better decision making [1]. IT and the health industry together are introducing new mechanisms for improving and monitoring the quality of patient care and quality of life.

Dementia is a key issue in health and social care given the context of an aging population. Dementia is "a decline in mental ability which affects memory, thinking, problem solving, concentration and perception" [2]. It is a disease caused by damage to the brain cells and is not specifically a part of the aging process and is progressive in nature with no cure. There are currently about 700,000 people with dementia living in UK and the number is expected to rise over 1 million by 2025 [3]. In 2001 there were an estimated 24.3 million people with dementia globally [3].

The quality of dementia care has been noted as an international area of concern [4]. Many countries including UK, France, USA, Norway, Sweden and Germany have been establishing national strategies and plans to deal with issues related to dementia and have placed dementia as a key research and practice priority [5], [6], [7], [8], [9]. The World Health Organisation report on an international consensus on policy for long term care for older people [10] specifically mentions quality assurance, research, data collection and strategic analysis among the key areas that international policies must address.

Assessing and improving the quality of formal dementia care is not a simple or short-term initiative. The needs of people with dementia are varied and often highly complex. Many people with dementia are unable to communicate their needs or experiences of care; therefore, systematic observation of care needs to take place to capture and attempt to understand the experiences of people with dementia and thus to raise the quality of care they receive. Dementia Care Mapping (DCM) [11], [12] was developed at the University of Bradford by the Bradford Dementia Group as a systematic observational framework and practice development process for assessing and improving dementia care quality. It involves continual observation of 5 to 8 people with dementia over a sustained period of time, usually 6 consecutive hours. Codes are recorded that represent the behaviour, mood and engagement level of the people with dementia, for every five-minute period. This is assessed by attempting to take the standpoint of the person with dementia. The quality of staff interactions with the people with dementia is also noted. Detailed process of the DCM system and how it works is described in Chapter 2.

DCM has been used as both a research and practice development tool internationally [13], [14], [15], [16], [17]. This research has shown that, if used over time, DCM can help to maintain or improve the well-being of people with dementia [13], [15], [18] and is a useful tool for research [18]. There are approximately 8000 people trained to use DCM tool (mappers) spanning over 5 continents. However, despite the success of DCM and the annual generation of a huge amount of data on dementia care quality, there is no mechanism or framework for large scale data collection, storage, sharing and usage. Currently data is collected and stored in Excel based spreadsheets and paper based files, which by no means provide the organisations using DCM a systematic way of saving, tracking and comparing data over time and to monitor progress or trends in care quality. DCM needs modern and sustainable approaches of IT contributions to facilitate a large number of national and international users to have access to quality data in a secure environment. Modern technologies and management systems together can provide highly skilled researchers, social and health care authorities and governmental authorities with a comprehensive framework to access the DCM data for analysis and decision making purposes. The DCM system lacks a comprehensive data governance framework which defines the structure of DCM data activities from capture to dissemination.

The rest of the chapter includes description on data governance in general and then its importance in the DCM domain. Motivation to take this study on board has been described as well. The contribution made in health care knowledge is also mentioned. At the end of the chapter the structure of the thesis is presented.

1.1 Data Governance

“Data Governance is a quality control discipline for assessing, managing, using, improving, monitoring, maintaining and protecting organisational information” [19].

Data Governance (DG) refers to the implication of policies and accountabilities to data management in an organisation [20]. The data management procedure includes availability, usability, quality, integrity, and security of the organisation's data according to their users and data requirements.

Every year a large amount of data is collected by organisations and it is expected that the information will be extracted from this raw data using different integrated and analytical technologies. This is a complex and time consuming process. Yet organisations spend a lot of money and time to accumulate this data and on its transformation into information. When integrating this data from disparate data sources, the need to maintain data quality, its security, access rights and its ownership initiates complex issues which need to be addressed to have a sustainable and a consistent data management process. DG is a process where the problems and issues of any organisation's data acquisition, handling, quality, security, access and decision making are addressed [21]. Different roles and responsibilities are assigned to process the data in an organisation. DG strategy helps to deliver appropriate data to authorized users when they require it [22]. A governance process in any organisation ensures the access of quality data to authorized users for research, analysis and decision making purposes [23].

1.2 Data Governance in Dementia Care Mapping

DCM is a process to evaluate the quality of dementia care in formal and informal dementia care settings across the world. Trained mappers collect data from different dementia care settings regularly. This collected data is processed and reported back to the dementia care settings for evaluation of the quality of care and quality of life in people with dementia. There is no method to-date to combine the DCM data nationally and internationally for comparison, analysis and benchmarking. Researchers do not have access to quality and secure data on DCM to explore the research issues generated in dementia and even in DCM itself. DCM data needs standard definition, secure access points by authorized users, monitoring of data quality and security, and dissemination levels. A data governance process will address all these requirements by implementing a comprehensive model. DCM data management process under data governance framework will facilitate the availability of the rich data to potential users under acceptable secure conditions, compatible with patient's data privacy issues, data quality issues and user defined requirements.

The quality of health care of people with dementia can be improved if the users have access to the DCM data for analysis, comparisons and decision making. This can only be possible if a data governance process involving efficient and sustainable IT methods is available to manage the data internationally. The data governance framework provides a comprehensive solution to manage the DCM, using a data warehousing approach, across the world from data capture to dissemination. Data warehousing [51] is a data management approach which provides data storage solutions for multidimensional analysis and decision making purposes. The proposed data governance model describing the DCM data management framework is discussed in detail in Chapter 4.

1.3 Aims and Objectives

The DCM system lacks a data governance framework providing an efficient and sustainable data management approach. This management system can improve the quality of health care in people with dementia, nationally and internationally, by allowing a variety of users from social and health care to use the DCM data for analysis and decision making purposes. Researchers can make use of benchmarking the DCM data to carry out predictive analysis and comparisons between different organisations' dementia care quality. To-date there has not been any attempt made to design a data governance framework describing a data management solution for DCM system [25]. The main aim in undertaking this novel study was to explore the steps needed to design a comprehensive framework for DCM data governance. This includes introducing and designing a DCM data management framework based on a proposed data warehousing approach to allow a wide range of users to have access to DCM data for effective monitoring of dementia care quality. Furthermore the aim was to present a novel approach of managing the DCM data from its capture to dissemination by applying existing IT techniques.

1.4 Contributions to Health Care Knowledge

This study proposes some significant contributions to the current health care knowledge in dementia. The first is a data governance framework design, to manage the DCM data from capture to dissemination, using existing methods of data management in health care settings. The second is designing a structure for a data warehouse to store the DCM data over time and defining accessibility and security models on this data, enabling secure data access by various users.

The original research work reported hereby will help in implementing successful strategies to increase the use of relevant knowledge in health care decision and policy-making processes.

1.4.1 Quality of Care

Improvements to the quality of dementia care can be supported by IT contributions to the DCM system through providing an efficient and sustainable solution to manage the data. A data governance framework to manage the DCM data from capture to dissemination using existing methods of data management in health care settings will provide a sustainable solution to capture, store and retrieve data. This will help a variety of users from social and health care to access good quality DCM data for comparisons, analysis and efficient decision making.

1.4.2 Dementia Care Monitoring in Different Dementia Care Settings

Having access to a historical and quality DCM data will enable dementia care providers in social and health care to monitor the care quality in different dementia care settings.

1.4.3 Standardization

The DCM data governance framework will provide the DCM users with standard guidelines to manage the data at different access levels. These guidelines will ensure the standardization of storage and access of the DCM data at national and international level.

1.4.4 Scalability

Scalability is the ability to expand any system's functionality to fulfil the growing amount of data or number of users without making major changes to the system [24]. The proposed DCM data warehouse structure for managing

DCM data will ensure the system scalability by providing a solution to manage a growing amount of DCM data over time.

1.4.5 Data Quality

Data quality is a very important issue in any kind of data used for decision making purposes and refers to a complete data [30]. This study will provide a governance framework which deals with this issue in a continuous manner. The data warehouse approach will provide a storage system which only stores quality and consistent data. Application of data mining techniques on the DCM data will help to identify the missing data, redundant data and attributes need to be collected for finding detailed information from raw data.

1.4.6 Data Security

Like any other health data, there can be several issues with security of the DCM data. Patient privacy is the main issue which needs to comply with data security regulations. A data anonymizing strategy can be a solution to secure patient's personal information at different access levels for secondary uses. DCM data will be collected from national and international data vendors and the exercise of data anonymization techniques will be different. The DCM data governance process will establish standard rules of data anonymization and data security.

1.5 Research Methodology

To design the DCM data governance framework a software engineering methodology was applied to capture the potential DCM user's requirements. DCM short courses were attended to learn about the DCM data and how the DCM system works. During these courses national and international DCM users were interviewed and asked questions to understand the end users requirements to develop the DCM data management stages e.g. DCM international database and DCM data warehouse.

Due to ethical and privacy issues with DCM data only limited data was obtained from some dementia care settings from UK. The acquired data was anonymized for the purpose of practical application.

1.6 Thesis Structure

The rest of the thesis is structured as follows:

Chapter 2:

Chapter 2 provides a detailed overview of the DCM system, how it is currently used, the users of the data and system. It also gives an overview of the existing techniques used in the DCM data processing and the need to manage the DCM data effectively.

Part of the material from this chapter is published in a paper at a renowned peer reviewed international conference [25].

Chapter 3:

This chapter presents relevant work in the field of health data management and comparisons with the proposed work in this thesis. This includes an extensive literature reviewed on different components of data governance. Existing systems of managing health care data across world have been discussed and criticised in detail.

Chapter 4:

The contributions towards designing a data governance framework for the DCM are presented in this chapter. Different DCM data governance components have been described as well.

Chapter 5:

This chapter explains the proposed framework of the DCM data management. Different steps to design the DCM data management have been described and discussed in detail. The proposed framework has been evaluated by using a limited DCM data. The small scale application results are shown with the implementation of the successful queries.

Parts of the material from this chapter were included in the publication [25].

Chapter 6:

This chapter provides an application scenario based on a case study to explain the application of data mining (clustering) on the DCM data obtained from the DCM data warehouse.

Chapter 7:

This chapter concludes the thesis by providing a summary of the thesis and a list of contributions made during this MPhil study. An overview of the open issues and the future work is also given in this chapter.

Chapter 2: Dementia Care Mapping (DCM) System

2.1 Introduction

This chapter provides a descriptive view of the Dementia Care Mapping (DCM) and how this system works. At the end, the limitations of currently applied data processing techniques have been discussed and the importance of a data management system for the DCM system is recognised giving some examples.

2.2 Background of Dementia Care Mapping

DCM [12] is an observational tool which was designed and developed by the late Professor Tom Kitwood, and Kathleen Bredin in the late 1980's. The purpose behind it was to develop a tool which assesses quality of care, in dementia care settings, from the perspective of people with dementia.

In 1992, the Bradford Dementia Group at the University of Bradford commenced training people, from different professional backgrounds, on the use of DCM tool. Since then this tool has been used in formal dementia care settings such as hospitals, care homes and day care.

According to Brooker and Surr [26] "*DCM is both a tool and a process. The tool is the observation and the coding frames. This is the intensive in-depth, real time observations over a number of hours of people with dementia living in formal care settings. The process is the use of DCM as the driver for the development of person-centred care practice including careful preparation of staff and management teams' feedback of the results of the map, action* *planning by the staff team on the basis of this feedback, the monitoring of progress over time and then the cycle of re-mapping commences"*.

The development of the DCM tool was based on identifying a process to improve the quality of life and quality of care for people with dementia. This tool works properly only if used carefully to record the sensitive data about people with dementia. This sensitive data is recorded by observing the behaviour and activities of a person with dementia over a particular time period.

DCM provides detailed and well structured information about well being and ill being of an individual or a group with dementia across a day. It provides information about how quality of care in care homes affects the quality of life of people with dementia.

2.3 How DCM System Works

The DCM process involves continual observation of 5 to 8 people (participants) with dementia over a sustained period of time, usually 6 consecutive hours [12]. Observations are carried out by a trained observer (mapper). Codes that represent the behaviours and mood and engagement levels of the participants for every five-minute period (time frame) are recorded. This is assessed by attempting to take the standpoint of the person with dementia. Two types of codes are recorded; Behaviour Category Code (BCC) and Mood and Engagement (ME) Value. The BCC represents one of 23 different domains of a participant's behaviour. The letters from A to Z (except H, M, Z) represent the main behaviour observed within a time frame e.g. A for Articulation (when a participant is verbally or non-verbally engaged in communication or interaction with another person or animal). ME values represent the mood and engagement of the participant associated with the BCC recorded in a time frame. The values of ME are expressed on a six point scale ranging from extreme distress (-5) to extreme positive mood and engagement (+5). So for example, a participant engaging in a positive conversation would be coded as A+3. Over a six-hour map up to 72 times frames of data may be coded for each participant. The quality of staff interactions with the people with dementia is also noted. These codes are called Personal Enhancers (PEs) and Personal Detractions (PDs). There are 17 different types of PDs and PEs that may be coded as and when they occur. All this information is recorded manually on paper-based raw data sheets and then part of this data is transferred onto Excel based spreadsheets for storage and basic analysis purposes. This information, in the form of reports, is then fed back to staff who develop action plans for care improvement. Observation takes places at regular intervals (3-12 monthly) to monitor progress and set new targets.

Currently, when undertaking data analysis, the mapper, for example, undertakes basic calculations such as percentage of time spent in each BCC and ME, and total number of PDs and PEs. They may also undertake further calculations such as calculating the average of all the ME values (known as the well or ill-being or WIB score), which represents the average level of well or ill-being experienced during the whole map by an individual participant or the group as a whole. The richness of the DCM data also permits other analyses to be completed for each map such as agitation and distress levels, withdrawn behaviour, passive engagement, opportunities for activity and engagement. This is done by combining different elements of the raw data into relevant categories.

2.4 Limitations of Existing Data Processing Techniques

DCM data is currently being stored in an unstructured, inconsistent and unlinked format (Excel based spreadsheets and paper based files).The data recorded in one mapping session is stored in different formats i.e. BCC and ME on spreadsheets and PD's and PE's on paper based files. Service user's basic details i.e. full name, age, gender and address are not recorded (even if recorded, for specific purposes, they cannot be related to the other recorded data about a mapping session). Lack of data management originates the integration complexities between disparate data structures. DCM data faces irregularities, incompleteness, storage, and quality issues. These data management problems currently make the DCM data unsuitable for complex analysis and recognition of trends and patterns over time.

In addition, researchers and practitioners are unable to get a broad view of a large amount of historical and consistent data, which could assist with monitoring and benchmarking dementia care quality locally, nationally and internationally for decision making authorities i.e. managers, executives. Government bodies are restricted in their ability to utilise the DCM data in decision making because of the lack of a structured, organised and efficient Decision Support System (DSS) [27]. Organisations have to go through a time consuming and costly process to bring all data together in one place to carry out very basic analysis.

This lack of data organisation and management inhibits the ability of the mapper and other potential users of DCM to utilise and analyse the DCM data to its full potential. The following queries, which can be extremely useful in care quality tracking and benchmarking cannot be answered:

- What is the average WIB score of all male and female service users in all UK units, each year, over the last 5 years?
- What is the average number of personal enhancers recorded per hour of mapping in Bradford unit, from 1995 to 2000?

2.5 Importance of DCM Data Management System

A huge amount of DCM data is generated every year nationally and internationally. This complex and rich data needs a sustainable and consistent data management framework [25]. The data management framework can facilitate the data storage, analysis, and efficient retrieval processes. DCM can provide the information of value to a variety of aspects of care delivery, which can help to achieve benefits on clinical, individual, group and organisational levels [28].

A data governance framework with a DCM data management structure will assist health care providers in monitoring and improving care quality more proficiently. Collected alongside other information about persons with dementia (e.g. diagnosis, severity of dementia, dependency, age, gender, ethnicity) and about the care setting (e.g. location, type, size, staff ratios), a single data repository of the DCM data will provide a picture of the quality of dementia care internationally, nationally, regionally and locally. National and international level storage and comparison or analysis of data over time, if available, would permit national benchmarking of care quality to assist in identification of DCM care quality indices, international comparison and tracking of care quality and also highlighting areas of poor and best practice. It would also provides an invaluable resource to researchers in dementia care, as it can be used in the longer term for quality of life related dementia research and for information retrieval.

My approach in introducing the data warehousing methodology to manage the DCM data will enable a variety of users (e.g. mapper, manager, researcher) to extract information relevant to them, from a historical data repository as discussed in Chapter 5. They will be able to store and retrieve the data in an effective manner.

2.6 Conclusions

This chapter explains the background and detailed introduction of the DCM system and how it works. The lack of a proper data management system in DCM has inhibited the variety of users from social and health care to use the DCM data for analysis and decision making purposes to improve the quality of health care in dementia care settings and quality of life in people with dementia.

Chapter 3: Literature Review

3.1 Introduction

This chapter presents the literature review on different components of the Data Governance (DG) process. The majority of this chapter is based on the work reviewed on data management systems designed to provide effective health care data management solutions, through the exchange of patient's medical data between different health care providers. The other important components of DG are data quality and data security, which are also discussed in terms of health related organisations.

3.2 Data Governance in Health Care

In the NHS and social care sectors, Information Technology (IT) based solutions to data collection and storage are becoming a central part of record keeping and monitoring of patient data. The governance of data required for such IT solutions has initiated the implementation of policies on data management, its quality, security and proper usage in a systematic manner [29]. Every organisation has its own way of governing the data, which depends on the organisation's data, policies, business requirements and areas need to be governed.

Very limited academic research has been done so far on the issue of DG. Most of the academic research related to DG is on the management of the quality of data [93], [30] which comprises only one part of the DG process. For the DCM data I will discuss the whole DG process which deals with all the issues of data management. Different organisation's including health care organisations are trying to deal with their data management issues through the implementation of a specific DG process.

The National Health Service (NHS) [38] of England has also moved towards IT based DG and implementation standards of data management, quality and guidelines of data anonymization, and use of data over time.

3.3 Data Management Systems in Health Care

Information Technology (IT) plays an important role in improving health care quality globally. Different IT applications have been applied by health sectors to manage the health data which is usually found in a disparate, complex and heterogeneous format. Mostly health care providers have developed local data management systems which are in incompatible data formats [51]. The disparate data sources need integration into a common format to facilitate the analysis, pattern recognition and decision making process, which otherwise is a time consuming and complex procedure.

IT in health care has been a valuable step towards health care improvement. IT contribution in health care was initially limited to basic data storage applications and Medical Health Records (MHR). However other examples of use of IT in health include the following applications in health care systems: Computerized Provider Order Entry (CPOE), Clinical Decision Support System (CDSS) [31], Picture Archiving and Communications System (PACS), Radio Frequency Identification (RFID) [32], and Interoperability [33] in health care [34].

Health data about patients, their health related issues, and their physician's details is collected annually and stored in different forms and structures. The combination of health and technology has the potential to improve the ability to manage the large amount of data into a consistent and easy to access format.

Different approaches to share the health care information have been applied globally. Two, in particular, are important [36]. One is the hub-and-spoke repository architecture which is practised in the UK, Canada, USA and Norway where data collected from different data providers is collected in jurisdictionally coordinated repositories. This provides a variety of users with access of timely and consolidated data. Another architectural approach is point-to-point information exchange systems which are being practised in New-Zealand, Denmark and Australia. In this approach each data provider develops and maintains its own databases locally and then data is integrated when needed on request. Both approaches are based on the common objective of providing the health care providers with integrated and quality data at the point of care to improve the quality of health care.

Some other applied and working data management systems in different health care organisations globally have reviewed and discussed. The details are as follows:

3.3.1 Canada Health Infoway (CHI)

Canada Health Infoway [35] is a Canadian organisation, developed in 2001.

This organisation provides an integrated framework of IT applications for health care organisations. Its basic aim is to provide different health care providers with accurate patient information, enabling better decision support at the point of care. The development of this organisation was based on the goal to improve the health care system's collaboration with each other and encourage the use of Electronic Health Records (EHR) system across Canada. This also includes the integration of all provincial health information systems into national pan-Canadian health network by the end of 2010 to provide secure and integrated quality information to health care givers and access to patient related health data [36].

The basic architecture of implementing the Canadian health network was based on hub-and-spoke repository systems which have service based connection, to collect the data from disparate data sources. This will allow the users to have access to timely consolidated data at the point of care. The Canadian health infoway chooses a series of hub and spoke coordinated repository system to store the patient's information because of two main reasons; one is that the infrastructure of EHR across Canada was not very strong at that time which created the problem for care providers to access the consolidated data and other reason is that Canadian government had financial control over the health care sectors [51].

Canada health infoway provides an information structure based on Service Oriented Architecture (SOA) approach [37]. This structure provides standards for data sharing across different health care sectors. SOA approach plays an important role in integrating the disparate systems based on services which are independent of underlying platforms and programming languages. The main concept is illustrated in Figure 1 described by [37].

Different components of this approach are

- Application Frontend
- Service
- Service Repository
- Service bus (HIAL)

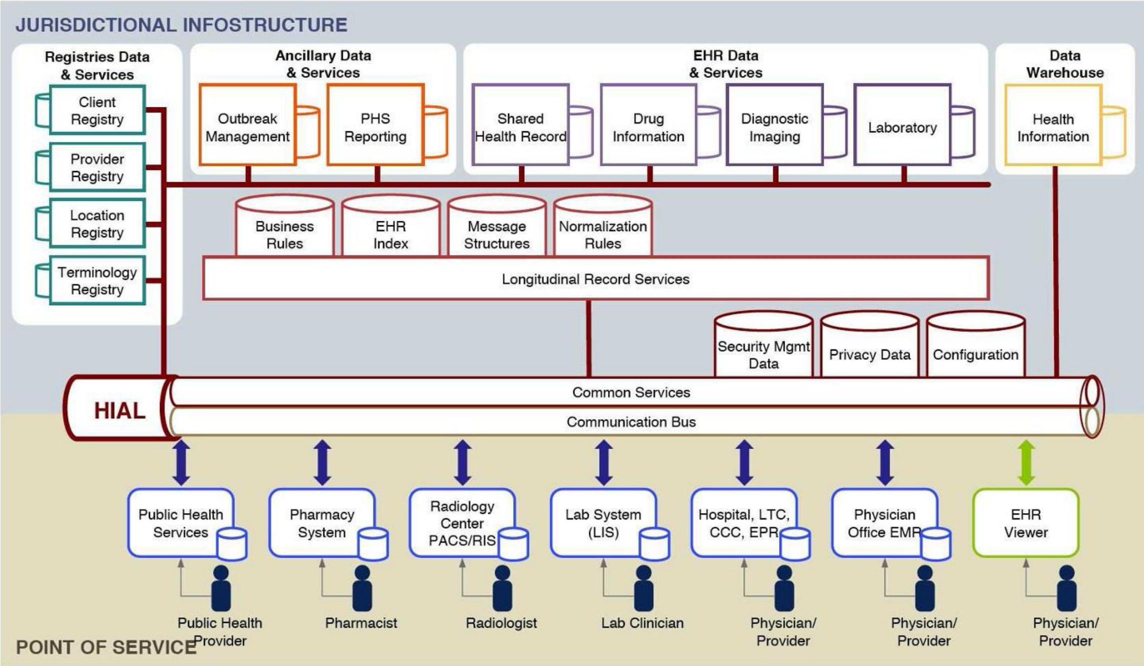


Figure 1: Canadian Health Infoway (CHI) Data Architecture [37]

Some additional features, for example, decision support systems and mined knowledge services have been introduced in this SOA architecture. These additional services will enable the health care providers to access the knowledge available in specific domain through standard services. Mining services will enable users to have access to quality and complete data for decision making purposes.

Canadian Health Infoway's IT network is based on the SOA architectural approach. This approach is based on the concept of supporting health information exchange nationally. The scalability of this approach has led the system network to expand by adding new components seamlessly. Adding decision support systems and data-mining systems into the existing network will be a challenging and complex action as the origin of the CHI was based on the idea of initializing the EHR nationwide and transporting the clinical documents through IT based communication network. But this system lacks the concept of automatic statistical analysis of patient data. This approach does not support the automatic historical data storage on atomic level, which could facilitate the analysis and mining on the data for future knowledge enhancement. Yet the extended SOA architecture opens new possibilities to add mined-knowledge services in the system but still the existing system needs changes to address the new added functionalities effectively.

3.3.2 NHS Centralized Summary Care Record Service

Connecting for Health (CFH) [38] is England's IT services provider agency working for National Health Service (NHS) in England. This agency has been established to provide IT based services to NHS for better and safer patient information storage and retrieval. The aim of this agency is to connect 100,000 doctors, 380,000 nurses and 50,000 other health care professionals and provide patients an opportunity to have access to their medical care records and health care providers to integrated data [39].

3.3.3 Spine

Spine [40] is an IT project established to create electronic health care records for more than 50 million of England's patients [38]. Spine is a national database, which is part of the NHS Care Record Service (CRS). It stores UK patient's electronic health care records summaries, which includes the patient's demographic and medical information.

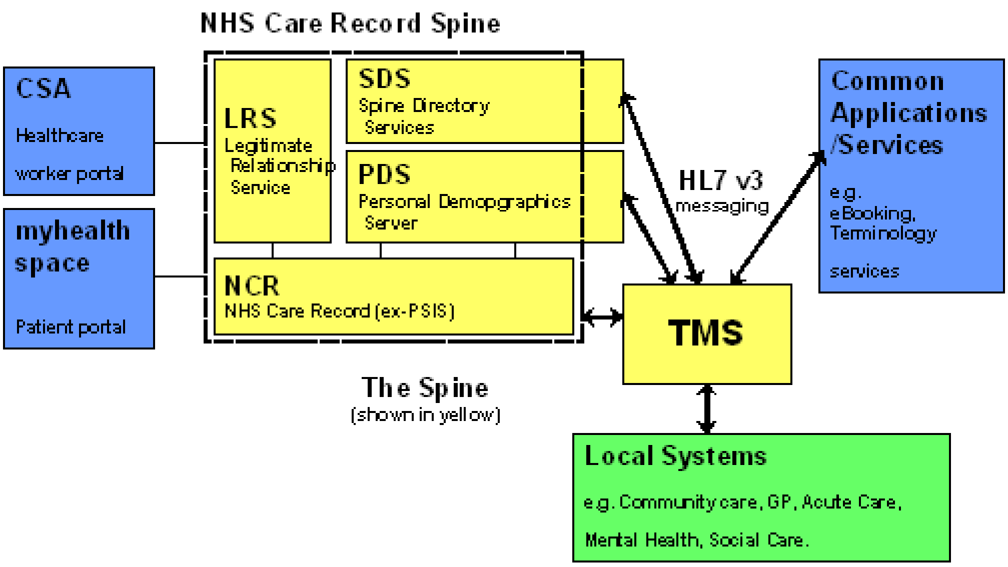


Figure 2: Spine Architecture by [39]

The goal of Spine is to gather data from multiple sources and deliver it to health professionals, who work in hospitals, primary care and community care services, for making decisions about patient's health care. The architecture of Spine given by Spronk [39] is shown in Figure 2. Spine consists of the following components:

Transaction and Messaging Spine (TMS)

TMS acts as a master "router" for messaging between all the spine systems. The transmitted message through TMS is based on HL7 V3 messaging standards [39].

Spine Directory Service (SDS)

SDS is an essential step for secure spine functions. This directory contains organisational details of GP practices. Each transaction and messaging service is monitored and only legitimate and authorized users can perform this action.

Legitimate Relation Service (LRS)

The function of LRS is to control the access rights on the patient's clinical data. This will enable different health care professionals to have access to consented data. Only authorized professionals are allowed to have access to patient's clinical and personal data.

Personal Demographic Service (PDS)

PDS is the central source of holding NHS patient's demographic information. This information includes patient's name, date of birth, address and NHS number. These patient's identifiers will be gathered from local databases [51]. Through NCR, service role based access will be given to a variety of users to protect the patient's privacy.

National Care Record (NCR)

Summarized clinical records are stored in this central repository. This repository collects the detailed data from heterogeneous sources and summarizes it for the use of different users for analysis and querying purposes.

There are different applications of this database; one of them is Secondary Uses Services (SUS).

Secondary Uses Service (SUS)

The SUS [41] is a data repository with anonymized patient records. This system helps researchers, analysts, practitioners and governing bodies to look at public health trends, medication and treatment efficacy, numbers of staff required by the NHS for future planning and other health care issues requiring planning or problem solving.

Summary

The described system architecture seems to be compatible with the user requirements. The basic aim of this system is to provide health care professionals with summary care records taken from various health providers. These records consist of integrated, summarized and anonymous data which is used for analysis and research purposes. But still there has not been attempts made so far to develop data mining technologies on the collected data to enhance the decision making and knowledge discovery process.

3.3.4 Biopattern Grid Project in Health Care

Biopattern, [42] is a European funded project. It is designed to provide analysis facilities on distributed bioprofile databases of people which will be remotely accessible via the internet to patients and clinicians. Biopattern is a pattern based on information, which helps clinicians to find evidences for diagnosis and treatment of diseases [42]. A bioprofile is a personal “fingerprint” that fuses together the person’s current and past medical history, biopattern and prognosis. The basic aim of this project is to make bioprofile data available from distributed databases for information sharing in a secure way via the internet. The other aim is to provide online algorithms, libraries and processing facilities [43] to support analysis on biopattren and bioprofiles to combat major diseases such as cancer and dementia.

The biopattern grid architecture described by [43] consists of four layers as shown in Figure 3.

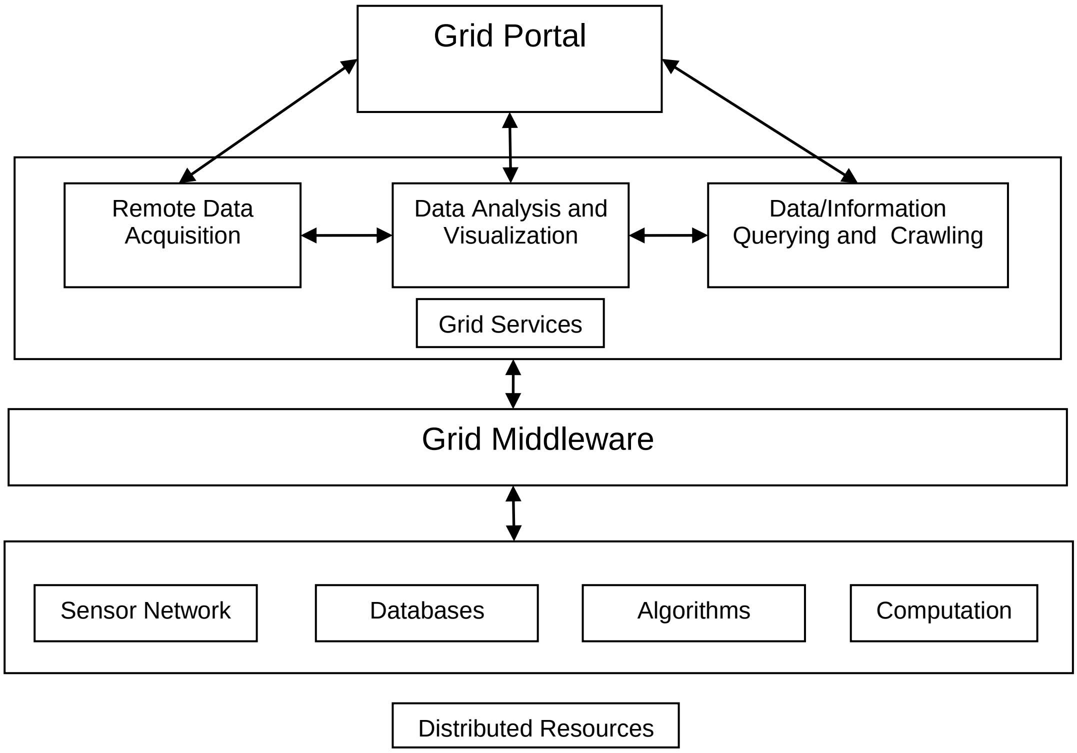


Figure 3: Biopattern Grid Architecture [43]

Grid Portal serves as an interface between end users and biopattern grid. Authenticated access is allowed by end user to access the services of BIOPATTERN grid.

Grid Services layer provide services for data acquisition, analysis, visualization and information querying and crawling. Grid Middleware provides functionalities of security, data management, resources management, information service and data service support.

The data distributed resources layer consists of different computational resources for grid functionalities i.e. data resources (relational databases), knowledge resources (algorithms for intelligent computation) and networks (sensor networks for data acquisition).

3.3.5 Bioprofiling over Grid for Early Detection of Dementia

The bioprofiling grid architecture has been applied to early dementia detection [44]. Several objective methods are available for potential early detection of dementia disease; one of them is EEG which measures electrical activities of brain. Bioprofiling grid approach is different from the early detection methods in a way that it provides comparisons between individualized care through subject specific bio data for analysis [43] not group comparisons. The idea of individualized care means that the individual patient's bioprofile data will be anlysed and will be compared over time to see trends and patterns in patient's previous and current condition[45]. For this purpose biopattern grid is used in order to analyse patient's EEG store in heterogeneous databases overtime. Clinicians are able to query, upload, update and analyse data through grid portal via web browser as shown in Figure 4.

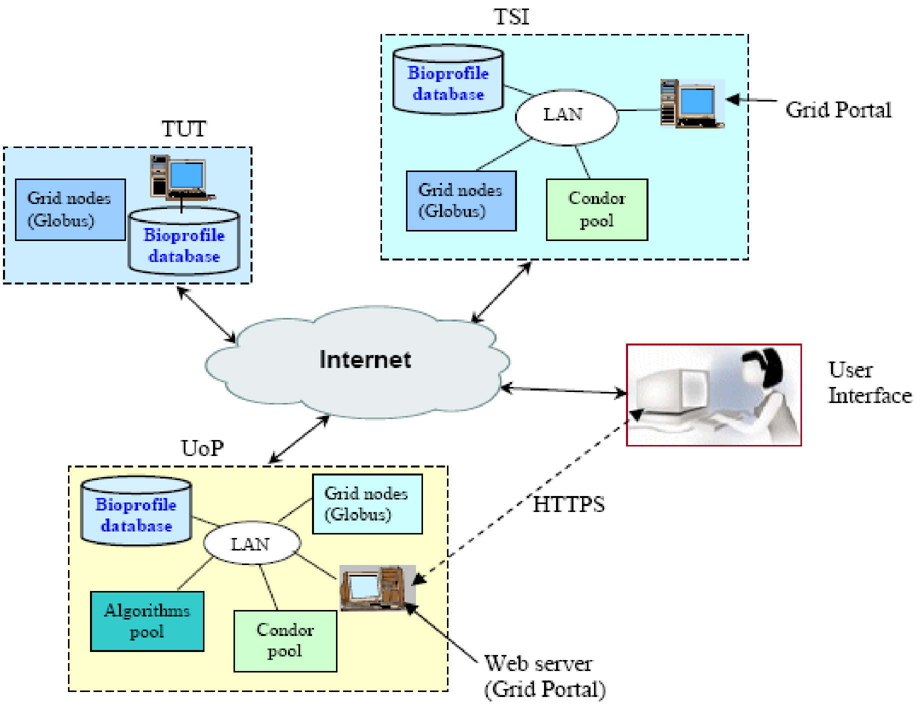


Figure 4: BIOPATTERN Grid Prototype for Early Detection of Dementia by [44]

Summary

Biopattern grid approach for interoperability in health care is demonstrated by using it in potential early detection of dementia. This approach provides analysis facilities on heterogeneous bioprofile databases to detect the patterns in data using different computational algorithms. But there are still security, integrity and ethical issues which need addressing. The mentioned approach has the downfall of not communicating and sharing data. Strong, reliable, secure and quality service needs to be implemented in order to enable seamless exchange of patient’s sensitive data between heterogeneous data sources.

3.3.6 Hospital Episode Statistics

Hospital Episode Statistics (HES) [46], is a national statistical data warehouse for NHS care provided to people living in England and those registered with the NHS but who are not living in England. The HES repository provides the data for analysis on an individual level and organisational level to NHS authorized personnel’s.

3.3.7 Bioterrorism Surveillance System

Data warehouse systems are also used in other aspects of health care and related areas. For example, in [47] authors describe the importance and contribution of a data warehouse in a bioterrorism surveillance system to recognise the patterns in medical data to identify abnormal situations and facilitate the analytical procedures to handle these situations.

3.3.8 Disease Management Programmes

In [48], the usefulness of a data warehouse approach in disease management programmes to help to maintain the health of the general population is outlined. The role of a data warehouse for hospital infection control is discussed by [49]. This paper describes the importance of a clinical data warehouse for research, quality importance, organisation and accessibility of data. They also highlight the importance of a clinical data warehouse in detecting medical measures and describe how this system can help to save money and time to acquire the precision in decision making.

3.3.9 Comprehensive Assessment for Tracking Community Health (CATCH)

In [50], the contribution of a data warehouse in providing the data for analytical and decision making purposes for Comprehensive Assessment for Tracking Community Health (CATCH) is discussed. The authors assert that a data warehouse will be a most demanding future application of technology in health care sectors.

3.3.10 An Evidence- based Health Care System

A data warehouse solution for decision-based health care has also been proposed and justified in [51]. According to Stolba et al, a data warehouse is a suitable solution for those medical systems which need decision support systems to enhance the quality of health care.

3.3.11 Assessment of the Described Systems

The work surveyed above highlights the need for integrating medical data present in disparate formats for sharing, analysis and decision making. The data warehousing approach has been applied to manage health care data where data is integrated from different sources in one repository and used for complex analysis and decision making purposes.

This study emphasises the need for a comprehensive solution to store and retrieve the DCM data, and I believe that a data warehousing approach will solve this problem. My proposed approach is different from most mentioned existing solutions in health care in a way that I have distributed the load of data integration into two steps: the international database will integrate data from those sources providing data for basic analysis, for example data taken from different dementia care settings, old DCM spreadsheets and other DCM related documents, on the other hand DCMDW will accumulate the historic data from international database and other DCM organisations local databases (anonymous data) for complex analysis, reporting and decision making.

3.4 Quality of Care and Benchmarking in Health Care Settings

The health care sector is among the largest service industries in developed countries. This sector produces out puts that account for about 7% of GDP in European countries which is more than the financial and trade service sectors [52]. Productive capacity of the workforce and wellbeing of general population is influenced by the output produced by health care sectors, which directly have impact on economy. To gain and maintain an effective and efficient health care sector, efforts are made to develop suitable metrics to monitor the performance of the health care sector. Health care providers are recognising the need to learn from others experience and drawing lessons on how to finance, manage and organise health care to improve the health care performance [53]. Benchmarking is a process to achieve this objective. The quality of health care in health sectors can be improved by benchmarking the health data to find effective and good practices. Benchmarking provides a performance assessment framework [53] which helps the health care sector to find the performance indicators that seek to capture the different aspects of the health care.

Different health care organisations have developed their own benchmarking data strategies. In the UK, the NHS has developed some information services which are responsible for providing benchmarking data to other peer organisations for quality checking and evaluation and analysis purposes. Some of these are discussed here.

3.4.1 NHS Benchmarking Network

The NHS benchmarking network [54] was developed in 1996. This organisation provides a networking and benchmarking services to its member organisations. This benchmarking facility enables different health care organisations to compare their data for identifying the best practice and areas which needs improvement. This improves quality of the health care, patient experience, productivity and effectiveness across different organisations. Data contributors are provided with a standard data collection template, regular comparison reports and summaries of key learning points. This system helps member organisations to have access to other organisations data as well to see and compare the analysed data. The NHS benchmarking network is working on different projects. Some of them are:

- Benchmarking long term condition
- Benchmarking mental health services
- Benchmarking shared and support services etc

According to NHS Midlands [55] benchmarking enables practitioners and providers to know what are the best practices and how can a change in process be introduced to deliver it. NHS Midlands introduced a toolkit “The Essence of Care” [55] in 2001 to help health care providers to recognise best practice by comparing and sharing data across different organisations. This benchmarking process was initiated to improve the quality of health care by ensuring consistency in quality care provision.

3.4.2 Benchmarking NHS Health Informatics Service

Health Informatics Benchmarking Club (HiBC) [56] was launched in 2008 by NHS Connecting for Health. This service enables health informatics services and IM & T departments to evaluate and benchmark their services with other member organisations to identify the areas of development and improvement. Member organisations can have access to each other’s anonymous data for identifying the comparison and analysis levels.

3.4.3 NHS Indicator Explorer

Indicator Explorer (IE) [57] is another service provided by NHS to provide comparison indicators for different NHS organisations for comparing their performance against national average trend rates. It also provides admission ratio indicators for Primary Care Trust (PCT) to find out the higher and lower confidence limits.

3.4.4 NHS Comparators

NHS Comparators [58] is a free analytical service which enables commissioners and health care providers to have access to benchmarking and comparing activity to improve the quality of health care delivered by their organisations. The health care organisations submit their different types of data which is compared and analysed with other organisations data to investigate the best practice and areas for improvement. For example the following data is collected by this service:

- Quality and cost data from Secondary Uses Services (SUS)
- Quality and Outcomes Framework Information (QOF)
- GP practice patient's demographic and prescribing data

Health Commissioners use NHS comparators to find out the referrals and access rates to secondary care in terms of cost and activity. This service allow different users to access an aggregate level of comparative access and performance rates by GP practice, PCT level, provider level or above.

3.4.5 OECD Health Care Quality Indicator (HCQI)

This initiative has been funded by Commonwealth international working group on quality indicator initiative (CMF QI) [59]. The aim of these initiatives is to develop a common set of Quality Indicators (QI) for health data comparison cross-nationally.

3.4.6 Assessment

The basic purpose of reviewing the above mentioned systems was to identify the benchmarking services provided by other health care organisations and how they are using this service to identify best practices and quality of care among different health care organisations.

DCM intends to provide benchmarking DCM data for a variety of purposes. Health care and social care providers will be able to have access to benchmarking DCM data for quality checking, monitoring the quality of care provided by different dementia care settings and evaluation purposes. Researchers will be able to compare their own data with DCM benchmarking data to evaluate the analysis questions.

3.5 Data Security in Health Care

Any medical data about any individual is considered to be extremely sensitive data. This data needs protection and security and cannot be accessed by unauthorized individuals. It is proven that IT applications improve the efficiency, quality, and collaboration in health care systems but at the same time it creates security and data protection issues as well [60]. Health databases are being widely used for research purposes across the world [61]. If data is un-anonymised and used for research purposes there might be severe security risks involved. For example Berman [70] discussed these risks and how these can violate patient’s privacy, threat to life, risks for loss of data security.

Berman also discusses how medical data should be freely available to researchers for data mining purposes if it is anonymised properly so that no patient can be recognised and patients' privacy cannot be breached. Disguising the identification of patients from data can change the total context of the data [62], [63], [68].

Data anonymization is a complex task [66] and reflects the social, political and ethical issues surrounding the anonymization requirements. How the end data should look after anonymization process and who will be able to get what type of information from it, are the questions that should be considered according to each organisation's data accessibility and confidentiality requirements.

It is a crucial and important decision that what information should be anonymised and what should be retained because different information has different perspective to different analysis. According to Rock [62] if only participant, whose personal data is anonymised, is able to recognise himself from that data record then this is the limit of anononimity on data. It is critically discussed that if data anonymization has served its purpose of protecting participants data privacy by different scholars in [64], [65].

Different methods to secure patient's data can be adopted as alternatives to data anonymization, for example, consent from participant at the data collection [66]. If data will not ever be used for secondary purposes then taking consent from patients in advance is meaningless and can reduce privacy protection [67] issues.

Other different alternatives of medical data anonymization have been discussed by [64] where he mentions methods to avoid anonymizing data which are:

- trusting researcher by making him sign a contract that he will not breech the patient privacy;
- taking consent from patients directly at the time of data collection;
- by giving original researcher a degree of control over the access and use of data;
- establishing an inter-research trust zone;
- with the passage of time, the data used for secondary use may be history and forgotten or ignored.

Corti et al [68] suggest removing the identifier and replacing them with pseudo anonymous data so that the meaning or context of data even after anonymisation does not totally disappear.

3.5.1 SPIN, a Real-life Example of Data Mining

The Shared Pathology Informatics Network (SPIN) [69] is a research project funded by US National Cancer Institute. The objective of this project is to develop computational techniques to allow researchers to query, via the internet, large virtual databases (including data from 20 medical institutions located in US) and find answers. SPIN will enable approved researcher to extract the anonymised patient's data records from any of the member medical institutes for analysis. In the SPIN network, the anonymised data about pathology specimen will be loaded, taken from member medical institutions, and researchers will be able to search their required human tissue specimen for their research. In this project the patient's data protection is the main priority and access to only anonymised data will reduce the risk of patient's privacy issues. SPIN follows the Health Insurance Portability and Accountability Act (HIPAA's) anonymization guidelines and standards.

Human subject research, where a researcher uses patient's records for their research, needs to follow security procedures to deal with such a sensitive data. Researchers do not need to have consent from individual patients for using information for their study but the researcher gets the anonymized data through a secure procedure or transfer to use for computational algorithms.

3.5.2 Anonymization Algorithms

Usually two types of processes are recognised to remove the patient's personal data from its diagnostic data which are anonymization and de-identification or coded process. Anonymization is a process where a patient's personal information or the information which can identify the individual (name, address, hospital no, date of birth and social security number etc) is removed and a false identification is given.

While de-identification process is different from anonymization process in the sense that patient's personal information is not permanently removed from the records but replaced with some suitable codes which are stored at a secure place and can be de-coded again to identify the patient. But with coded data there are some risks involved in violation of patient's privacy. For example if master codes are accessible to any un-authorized person he can decode the patient's personal information. Or other example is if data record contains person's date of birth, ethnicity and zip code, along with diagnostic information then anyone intending to violate the privacy can identify the patient using public records (national demographic data, birth records etc.) [70].

To-date a number of techniques and methods to anonymise or de-identify data have been proposed. Two of them are Naïve Approach and K-anonymization technique.

Samarati and Sweeney in [71] criticise the Naïve approach by mentioning that this approach is not a strong option to anonymise data as if the anonymized data set is joined with other public databases then other can re-identify the individuals from the data. They strongly favour the k-anonymity technique including generalization and suppression process to de-identify data sets. k-anonymity process provides data for public use with no identifiers included or any chance to re-identify by any joining methods.

There has been a great deal of research done on K-anonymization using generalization and suppression processes [72], [71], [73]. Polynomial time algorithm for optimal anonymity [74] clustering techniques [75], [76] on K-anonymization method to achieve better quality data set without loss of the important data [73] and full-domain generalization algorithm for anonymity [77] which gives short runtime are some examples for current data anonymization techniques being applied.

Not all anonymization techniques are reliable, some leaves datasets difficult to interpret, have low performance problems [78] and result in information loss. Applying data anonymization techniques in medical data [79] is an important and compulsory task. Machine learning approach for data anonymization by introducing iterative named entity recognition method has been discussed in [80] and assumed to be helpful in securing patient’s health data.

HIPAA [81] in US allows medical records to be shared and used by researchers and public without patient consent if the identifier data is removed from the data set (anonymisation) prior to use.

3.5.3 Assessment

Health data including patient's personal or identifiable data needs to be de-identified for use in analysis or research purposes. Very strict regulations and guidelines are being practiced to secure patient's privacy. Data anonymization and de-identifications are the procedures which can be applied to secure patient's data and still be usable for research purposes. Some scholars have discussed the alternative methods to maintain the patient's data security without changing the context of data through anonymization method. Every organisation adopts its own methods and strategies of securing patient's personal data according to the end user's need and analysis requirements. In DCM the data anonymization step is very important to provide the data for secondary uses. The application of data anonymization techniques in DCM is out of scope of this study and is described as the part of the future work.

3.6 Conclusions

The purpose of this literature study is to explain the different data components of data governance in an organisation. These components are data management, data quality and security. Different real health organisation's data management systems and their data architectures are reviewed and critically evaluated in the summaries given at the end of each section. Different health organisations are practising different methods to bring data together for analysis and sharing purposes. But the underlying concept to bring the data together to be available to a variety of users of systems is more or less same and reflects the organisation's users and analysis requirements. Most reviewed health organisations are using a data warehousing approach to integrate the data in a single repository for further analysis and decision making. Data quality and security also plays an important role in managing health data. The real systems, practicing and implementing different techniques to provide quality data for benchmarking and secure the data for research and other analytical purposes, have also been discussed.

This review can be concluded by mentioning that health related organisations do need sustainable data management systems to bring disparate data together into one place where the data is good and secure to use for further analysis, benchmarking, data mining and decision making purposes. DCM data comes under the medical data category and reflects the same problems and issues which any other data on human subjects and their diagnosis have. I propose to apply the data warehousing approach to manage the DCM data where the modifications are done according to the DCM data and user's requirements.

Chapter 4: Proposed DCM Data Governance Framework

4.1 Introduction

This chapter presents the proposed DCM data governance framework. The framework identifies the DCM organisational needs and areas needed to be managed. An overview of data governance processing in the health sector is provided below and identified stages of data governance for the DCM system are proposed and explained in this chapter.

4.2 Data Governance Need in Health Sectors

Health care organisations generate and hold large amount of data and have been facing challenges to store and retrieve this data. Over the last two centuries the use of data in health care organisations has changed [82]. These organisations need data management solutions for an effective usage of large amounts of data by a variety of users to improve the efficiency and quality of the health care. For this purpose a variety of innovative techniques and strategies are being applied.

To improve the quality of health care, the health care industry needs to take some important measures and actions [83]. One of these is the management of data by introducing IT applications in health care [84]. IT has been proved to enhance the overall quality and cost of health care [85]. Different countries are trying to improve the quality of their people health care by introducing IT based data management systems in this sector [86], [87], [88]. The need to manage disparate data effectively without compromising its quality, security and privacy using modern IT techniques has pointed the need for data governance in health care organisations. A Data Governance (DG) framework helps organisations to establish roles and responsibilities to manage data and define structure for data management based on the organisation's data analysis requirements [89]. In the health care sector, governance towards medical data identifies the issues and concerns for data managers to handle data according to the user's requirements. DG is a process which deals with data acquisition, storage and usage within an organisation in a secure and reliable environment for a variety of uses. In the NHS and social care sectors, IT based solutions for data collection and storage are becoming a central part of record keeping and monitoring of patient data. The governance of data required for such IT solutions has initiated the implementation of policies on data management, its quality, security and proper usage in a systematic manner [90].

4.3 A Contribution to Design a DCM Data Governance Framework

Data governance in any organisation is a process to analyse the organisation's needs, identify initial areas of focus and then based on these findings establish plans and implement strategies [91]. Every organisation determines its own process of DG, which reflects the organisations data, users, access levels and business analysis requirements [89]. For the DCM organisation we propose a specific DG framework describing different generic components and subcomponents of a data governing process, from data acquisition to usage. By introducing DG in the DCM, we will be introducing standards, rules and roles to manage the DCM data at various usage levels nationally and internationally. The proposed DCM data governance framework can help to accelerate the access to complete and quality data to a variety of users. This will help to enable more efficient and proactive care in health care settings providing dementia care across the world.

4.4 The Purpose of Designing the DCM Data Governance Framework

The main purpose of designing the DCM Data Governance (DG) is to define a management structure to acquire data from different national and international DCM organisations and the provision of short term and long term data storage solutions where data can be accessed by users for variety of purposes. The proposed DG framework identifies the main users and their roles. These users will have access to the DCM data at different security and anonymity levels.

Other aims of designing the DG framework are to implement methods to identify data irregularities and inconsistencies and exercise data privacy and security laws on the data according to the defined rules. Standardisation of data across DCM organisations is also an important aim for designing and proposing the DCM DG framework.

4.5 Basic Components of the DCM Data Governance

There are different levels or components of describing DG framework in any organisation [92]. According to the DCM organisational requirements I divide DCM DG framework into three basic components:

- DCM Data Management
- Data Quality
- Data Security

The above mentioned components are usually the main recognised data governing issues in any organisation [82]. The designed architecture of the DCM DG framework is shown in Figure 5.

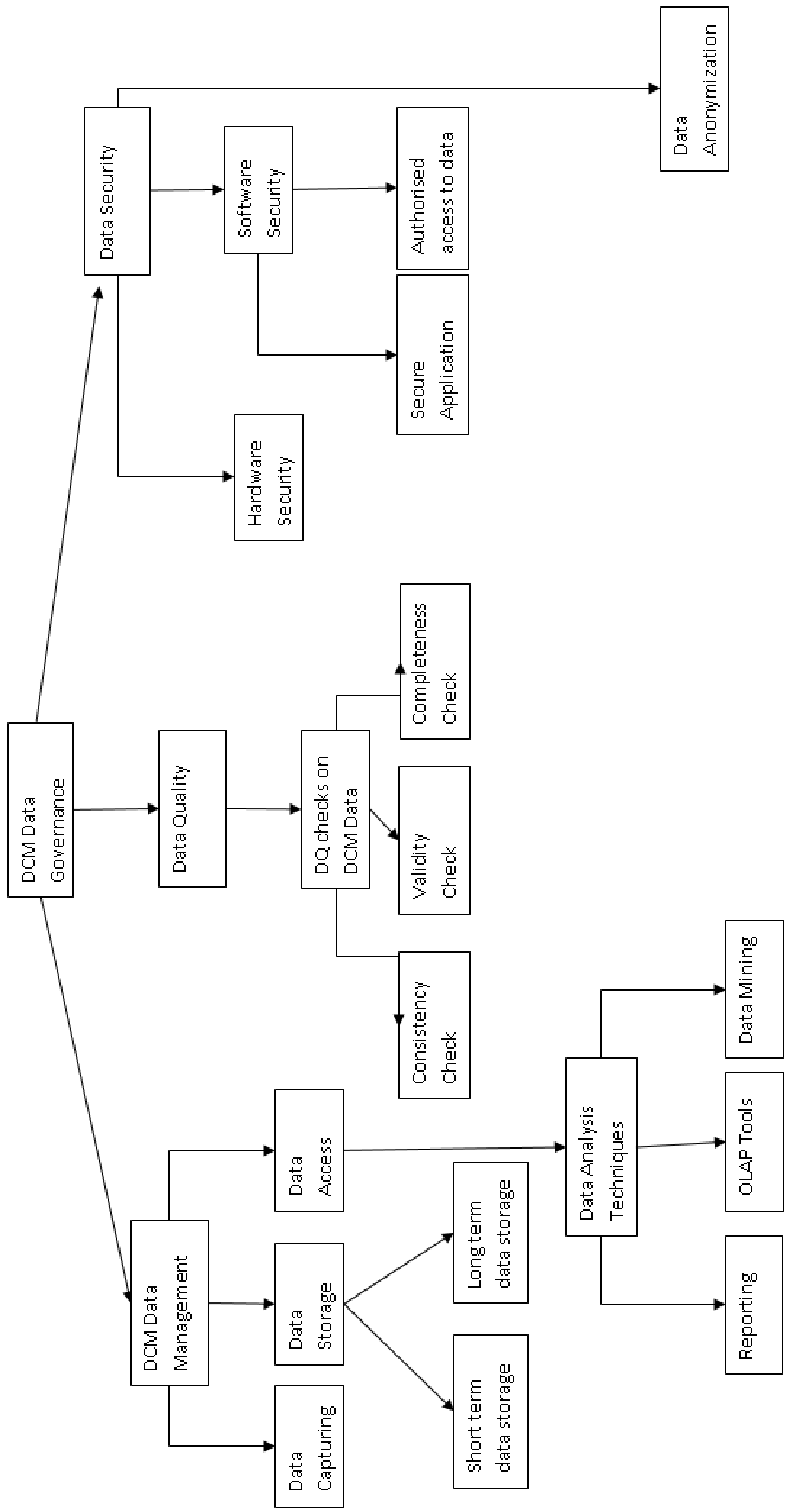


Figure 5: DCM Data Governance Framework

4.5.1 DCM Data Management

DCM is being stored in variety of formats and styles. Organisations have to go through a time consuming and costly process to integrate data at one place even for very basic analysis. Some data is in paper based files and other in spreadsheets. To perform analysis and decision making activities data needs to be integrated first into a consistent format. In any organisation the data management process reflects the end user's and organisation's needs. In the DCM organisation data needs management from capturing and integrating to dissemination. We divide the management process into three main stages

- Data acquisition stage
- Data storage stage
- Data retrieval stage

DCM data needs a systematic and efficient way to acquire the data from different national and international dementia organisations. The acquired data needs to be stored in a sustainable repository for further uses, for example basic analysis and even for long term storage for complex analysis, decision making and data mining. I identify the need to design data repositories for short term data storage for basic analysis and long term data storage for complex analysis and decision making. For this purpose I proposed a design of international database and data warehouse to store DCM data for different analysis purposes by variety of users. This contribution is explained in Chapter 5.

A variety of users need to access the DCM data at different anonymity levels for further analysis, comparisons, benchmarking, decision making and research purposes. These users belong to a variety of backgrounds including social care, health care, executives, practitioners, academics and governmental authorities. Figure 6 describes the DCM organisational structure showing a variety of users in a tree format. At the top executives and managers of organisations are shown who need to access DCM data at different granularity levels. In this figure the mapper role is not shown as this can be played by any member of the DCM organisation or any person who has successfully completed the DCM course organised by Bradford Dementia Group. Predominantly site managers, unit managers and unit staff are trained to be mappers. Mappers are responsible for collecting data from the dementia settings and feeding it back to the unit staff.

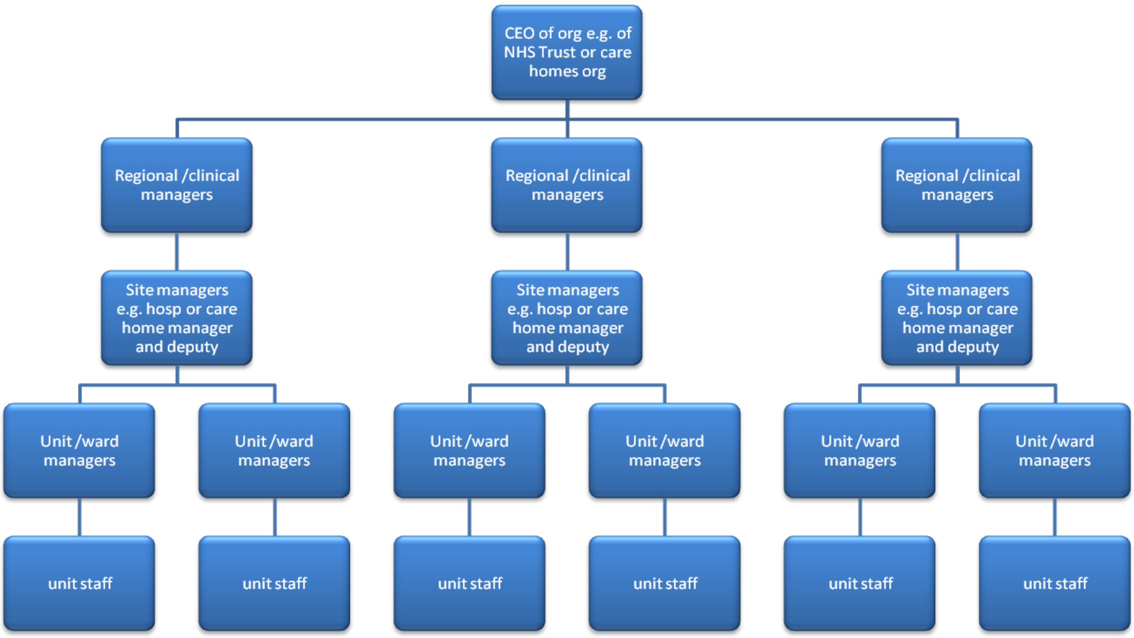


Figure 6: DCM Organisational Structure

Different access levels needs to be defined to manage the DCM data internationally. Data will be retrieved by a variety of users by using analytical applications i.e. reporting tools, Online Analytical Processing (OLAP) tools and Data Mining (DM) tools.

4.5.1.1 Data Access Levels in DCM

DCM data is rich data and can be used for certain purposes, for example, quality analysis, predictive analysis, research purposes and decision making.

For example a variety of users from social and health care are interested to have access to the DCM benchmarking data to compare the quality of care provided in different dementia care settings. Some end users, like practitioners, are also interested in predicting the levels of care needed based on the existing figures or measures on service user's well or ill being recorded in dementia care settings. Executives and managers can be interested in historical and quality DCM data to carry on decision making processes, for example, what and where and at which level the quality of care needs improvement.

DCM data needs to be accessed at different security levels throughout the organisation. These levels are shown in Figure 7. A mapper is a user who collects data from the field, analyses it and feeds back the results (reports) to the staff of a mapped dementia unit for further improvement in action planning and quality control. A mapper accesses the data at the highest granularity (where data is at its atomic level) and lowest anonymization level (mapper can identify the data about patients of his/her mapped units).

Managers need access to aggregated DCM data for general analysis purpose. For example they need answers to the questions like “What is the WIB score of all the dementia units including long term and short term dementia care homes in Yorkshire area in the year of 2009 and how does it compare with their own care homes?” or “What is the average score of well being of all male service users living in their care homes in year 2008 and 2009 and how the score can be increased in 2010?” The DCM data needs to be aggregated and anonymized to be accessed by unit managers and branch managers to answer these questions.

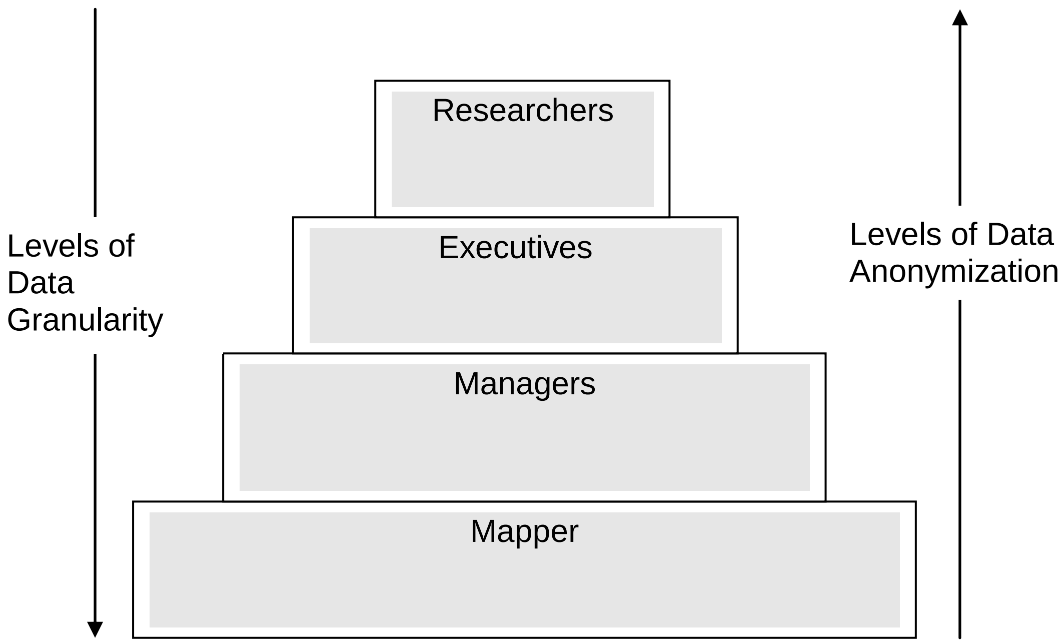


Figure 7: DCM Data Access Levels

At the top of the DCM data access hierarchy (as shown in Figure 7) the level of anonymization increases, the users e.g. executives (CEO’s) will be interested to have access to a highly aggregated data to know the averages for their organisation as a whole or for large organisations, region by region as well to compare the regions. These executives will be able to identify their own organisation's data but the data from other organisation will be anonymous to them. The following type of questions may be asked like "What is the agitation level of all service users living in dementia care homes in West Yorkshire in last 10 years and how does it be comparable with the agitation level of service users living in Scotland's care homes?".

Researchers will have access to only anonymous data (level of anonymity will be defined according to end users requirements) where they can compare and analyse the general anonymous data to carry out their research activities on the DCM and dementia.

4.5.2 Data Quality

Data quality is a major issue in any organisation. Data quality usually refers to the data accuracy [82]. Inaccurate and incomplete data leads to the wrong analysis results and poor decision making by executives and managers. Most of the DCM data is presented in the form of codes about service user's mood and behaviours and interactions with unit's staff members. These codes have specific meanings and are recorded by trained and experienced mappers according to pre-defined DCM rules. Critically the quality of analysis depends on the correctness of these codes. Furthermore, the integration of these codes with other data collected about service users e.g. personal details, diagnostic details is also an issue. Service users having specific coded values should be related to their correct demographic details or their correct diagnostic details to answer a variety of questions related to their health care and quality of care in dementia care settings. The proposed DCM data governance framework addresses these issues by providing data quality controls at the system data acquisition stage and during integration process as well. A data warehousing approach to manage the DCM data will solve these problems by providing a variety of sustainable methods to check the quality of the DCM codes, data correctness and completeness. The data quality issue in DCM has not been part of this study its inclusion in the DCM data management framework will be addressed in the future.

Wende [93] presents a data governance model based on Data Quality Management (DQM) of organisations where he discusses how to identify the roles and decision areas in an organisation to enhance the data quality of any organisation and methods to assign the responsibilities. According to his perspective, both the technical and organisational contribution is important to managing the data quality in any organisation.

4.5.3 Data Security

Medical data or data about human subjects is considered to be very sensitive. Privacy and security issues play a very important role in any organisation’s data governance or management process. Data protection laws vary from country to country. DCM data which will be collected internationally will be subject to such laws at source. DCM data security is not just limited to the service user data but it needs be a multifaceted security framework which encompasses the entire data governance process. The areas of concern will be security management policies and procedures, network architecture, hardware and software design and other critical protective measures.

Strong access control measures will need to be implemented in order to restrict access to service user data by DCM organisational roles. Physical access to data will also need to be restricted.

For the scope of this study we do not intend to discuss the DCM data security and anonymization techniques in detail.

4.6 Acquired Goals in the DCM Data Governance Framework

Following are the acquired goals and aims for DCM DG framework:

- Data acquisition standards and guidelines
- Sustainable data storage solution
- Authorized user roles
- Build standards and repeatable processes
- Interoperability between different health care organisations
- Enable better decision making
- Reduce operational frictions
- Achieve data security

4.7 Conclusions

The described DCM DG framework deals with the issues of data acquisition from a variety of users and sources under some specific rules and its dissemination and usage. The required DG components are identified. These components are divided into subcomponents to enhance the understanding of the DCM organisation and its needs. Data quality and security issues are part of managing the organisation's data but can be described and tackled individually if needed.

The scope of this study is limited to design the DCM data management framework by describing different methods of data capturing, storing and retrieving. The proposed structure of DCM data management is described in the next chapter.

Chapter 5: DCM Data Management Framework: A Data Warehousing Approach

5.1 Introduction

The methodology applied to solve the DCM data management issues through a data warehousing approach has been discussed in this chapter. Different steps have been identified while designing the DCM data management framework. All the steps with detailed examples have been presented in this chapter.

5.2 Proposed DCM Data Management Framework

As described in Chapter 3 there are different approaches and methods to manage the data in health care organisations. According to the DCM end user's requirements and data accessibility issues, I propose a data warehousing approach to manage the DCM data from capturing to dissemination.

5.3 Data Warehousing

Data warehousing is a process of combining data from multiple or varied sources into one comprehensive and easily manipulated database [118]. Different accessing systems or applications can be applied on the data warehouse data including queries, reporting, multidimensional analysis and data mining. According to Inmon "a data warehouse is a subject-oriented, integrated, time-variant, and non-volatile collection of data in support of management's decision making process" [94].

Data warehouse technology has been effectively applied into different industries i.e. retail [95], telecommunication, finance services, and travel [96] industries. This is a challenging and complex process in health care systems because of the complexity and largely unstructured formatting of medical data. The other reasons are the volume, complexity, and security of health data [97]. But in spite of these challenges attempts to develop data warehouses in different areas of health services have been made across the world and have proved to be successful to date [98].

A data warehouse is a data repository consisting of historical data taken from heterogeneous data sources. For managing the DCM data a complete data management system needs to be developed which deals with capturing data (taken from mappers, spreadsheets, flat files) into a data repository, cleaning and transforming and loading this data into a data warehouse and using the subject-oriented data from the data warehouses for different applications e.g. OLAP tools, data mining tools and reporting tools. I propose a DCM international database as a data repository (short term data storage) to store data taken from international dementia care settings practising DCM and from spreadsheets. This international database will act as a first point of data integration and storage platform. To collect the data from national and international DCM organisations a web application is needed. This web application will access the DCM international database for data retrieval as well as by mappers and researchers.

The DCM international database and other databases from other countries will provide data to be stored in the DCM data warehouse for multidimensional analysis. The architecture of our proposed DCM data management framework is shown in Figure 8. This framework explains different stages of data acquisition, storage and transformation.

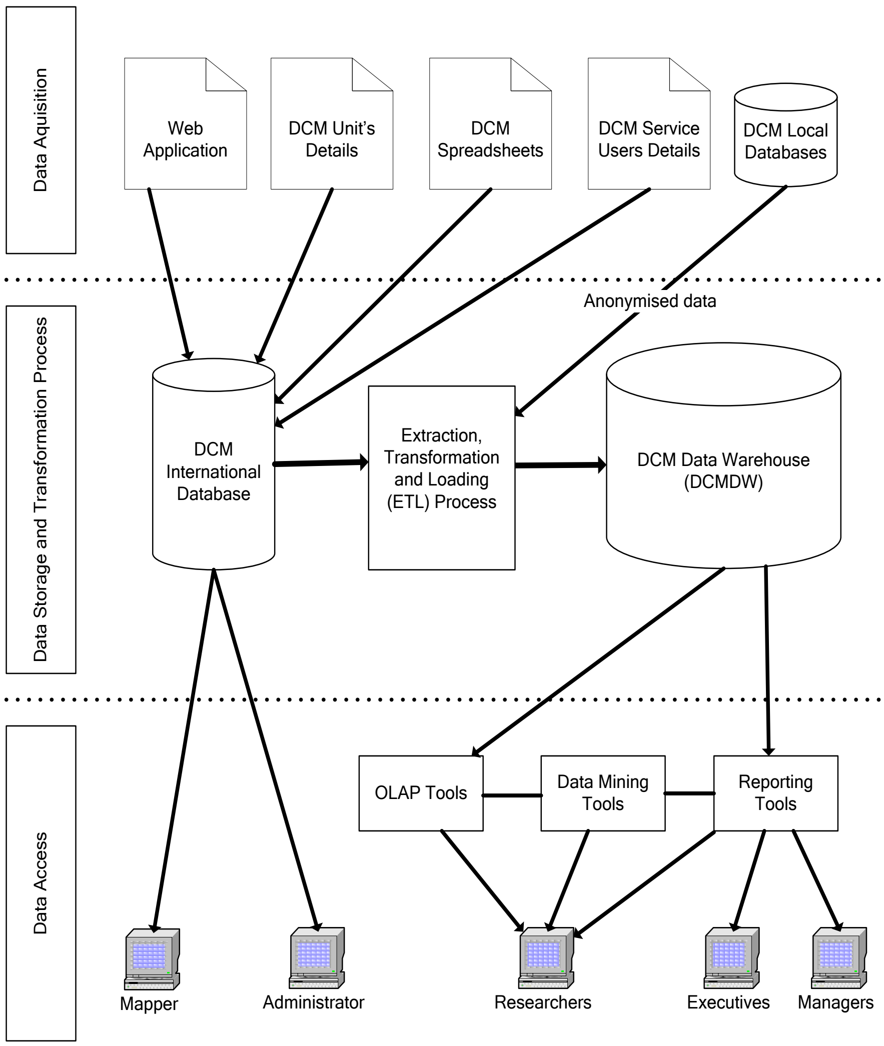


Figure 8: Proposed DCM Data Management Framework

5.4 Proposed Stages for DCM Data Management

As described earlier in Chapter 2, DCM data is currently stored in disparate formats. The potential DCM users are unable to access the large amount of data for further analysis and research purposes. The proposed approach of managing the DCM data [25] solves these problems by introducing data management through different stages which are shown in Table 1.

Table 1: DCM Data Management Stages

Stages	Applications	Data Source	Users	Benefits
Stage 1	DCM Relational Database(with web interface)	Different DCM organisations, old DCM spreadsheets, other DCM related documents.	Mapper, administrator	Best for data capturing, storing, validation and integrity. Global access.
Stage 2	Extraction, Transformation, Loading (ETL) process	DCM relational databases, Other databases(with anonymised data)		Data cleansing, integration and quality check
Stage 3	DCM Data Warehouse (DCMDW)	Different data sources	Data analysis tools, data mining tools	Easy data access and retrieval from dimensional structure of data
Stage4	Analysis Applications On DCMDW	DCM data warehouse	Managers, executives, governmental authorities etc.	Decision making, bench marking, analysis and reporting

5.4.1 DCM International Database

DCM data management framework describes different stages to manage the DCM data. The first stage is to develop relational databases for storing the DCM data and allowing mappers to do basic analysis on this data. This relational database will be accessible to store the DCM data taken from different national and international DCM organisations through a web-based standard interface. For DCM data capturing I develop an application based on web technologies (screen shots are shown in appendix D). This step assures the quality data entry into the database from the very first stage. Data can be retrieved by users for basic analysis, for example mappers can retrieve the calculated measures of stored BCC's and ME's.

For the purpose of designing the international data repository for DCM data a relational database is designed. Different entities and their relationships have been recognised from the DCM system. Following are the main identified entities:

Mapper, Service User, Participant, Unit, Mapping Session, Data Item, DCM User, User Roles

The relationships between these entities are shown in DCM Entity Object Model in Figure 9.

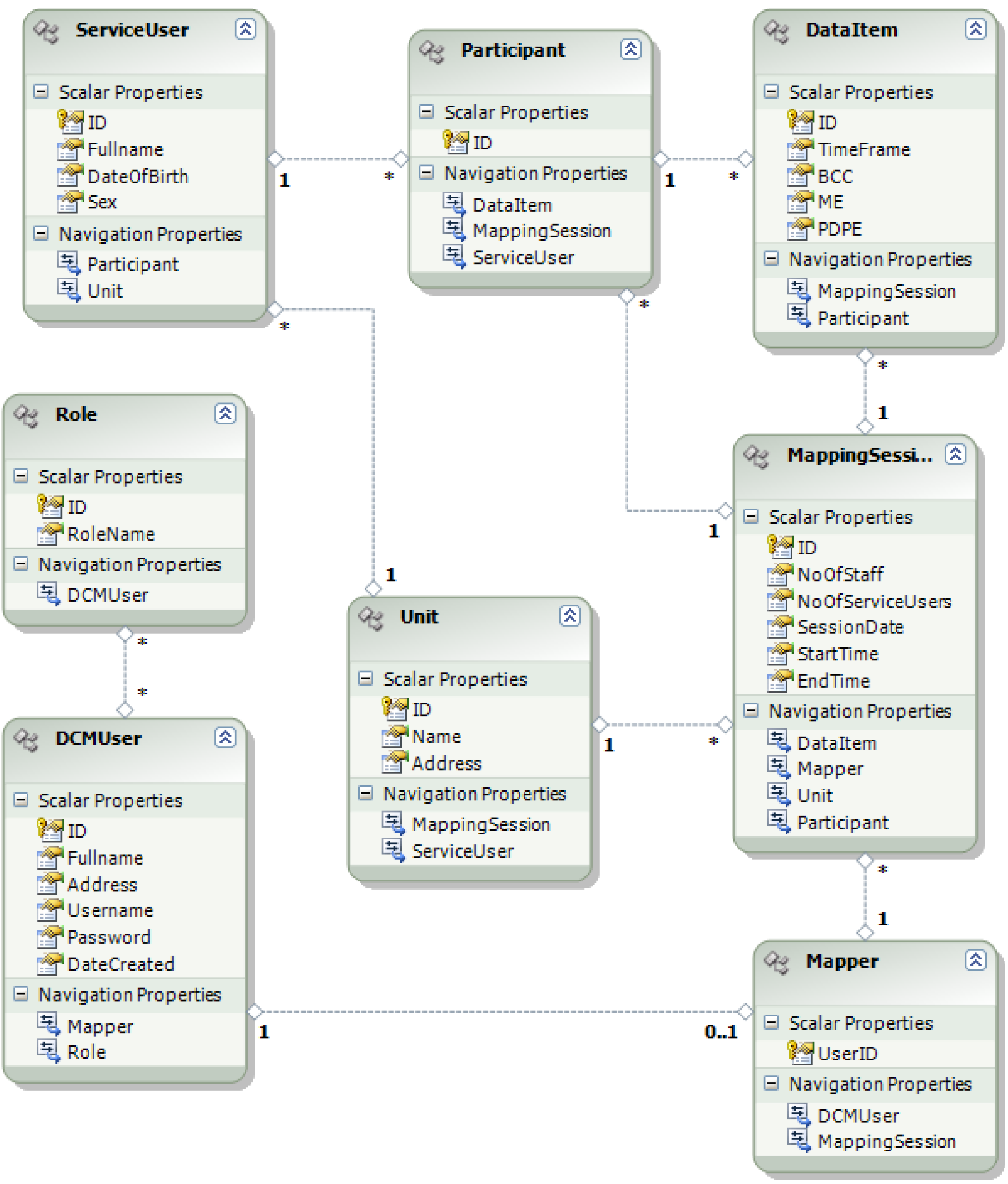


Figure 9: DCM Entity Object Model

DCM spreadsheets were in zero normal form (un-normalized) with data integrity and redundancy issues. These DCM spreadsheets were transformed from zero normal form to third normal form to bring the data into a suitable structure which is good for querying purposes and free from update, insertion and deletion anomalies [99]. Normalization is a method for ensuring that the data model meets the objectives of accuracy, consistency, simplicity, non redundancy, and stability [100]. For an effective and efficient relational database model, the attributes should belong to a proper entity and each entity should have an identifier (primary key) and for cross reference with the other object or entities the foreign key [101]. During the normalization step data inconsistencies were removed from the data to bring it into a normalized form.

The recognised DCM entities and their relations with each other are shown in Figure 10 as a logical schema for the DCM international database.

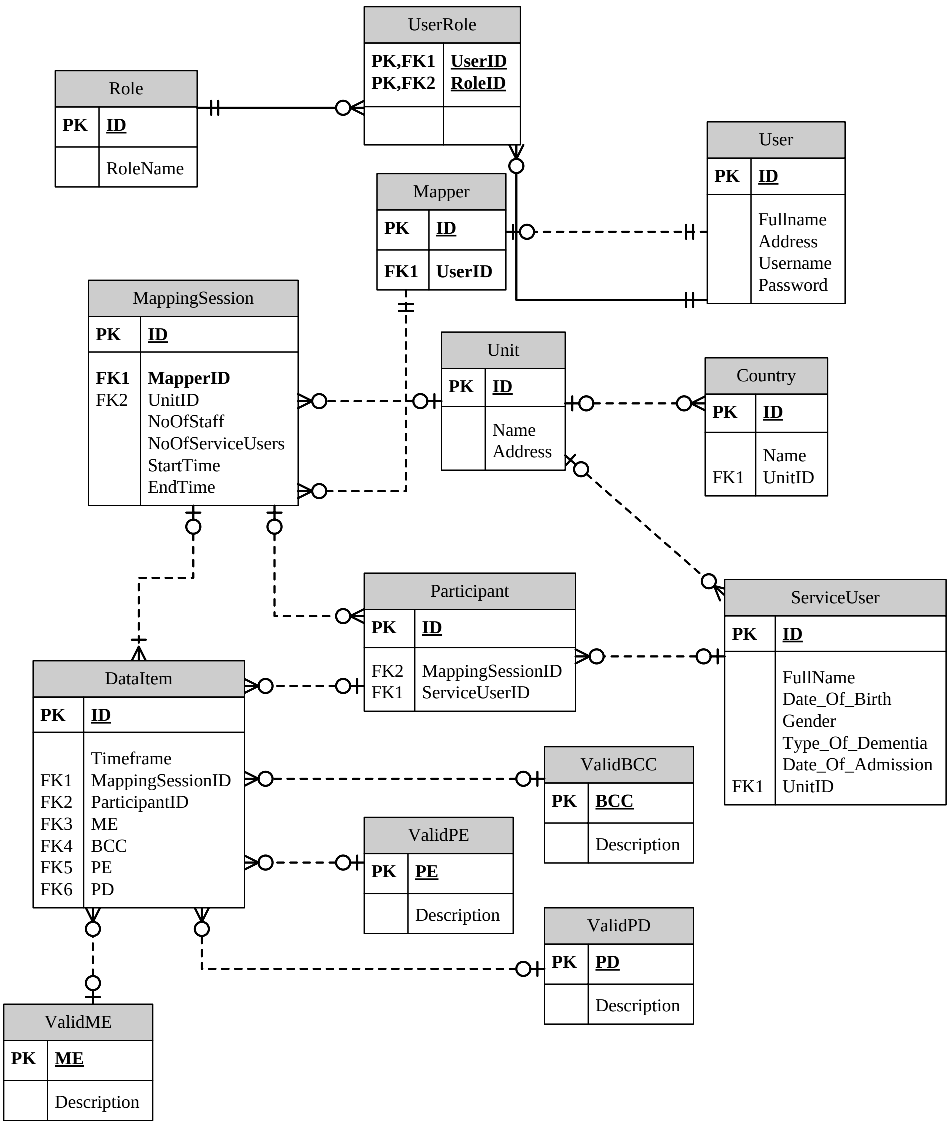


Figure 10: DCM Entity Relationship Model

Different roles (users) are identified while gathering the requirements. Each role can have access to the particular data to which it is authorized. In the designed and developed DCM application the mapper can only see the data of a unit that he has mapped for specific time period. The data accessed by other users e.g. researcher is pseudo-anonymous (service users are given unique identification). Mappers are able to use the data for basic calculations which are done with pre-defined queries (usually what is required from the mapping where the user is able to extract the reports to feed back to the unit staff). This database will provide a platform to integrate the data, taken from different DCM data providers and from other sources (old spreadsheets and paper based files), at the very first stage.The rule-based web interface has been developed to allow only quality data (validated DCM data) to be stored. This is the first stage where data quality is checked. There are different BCC’s, for example H, M, and Z, which are not part of the coding system and some BCC’s e.g. C (withdrawn behaviour) can only have an ME of -1 associated with it. These validations are done in database design. This procedure will ensure the correct entry of BCC’s and ME’s in the database. Old data (spreadsheets) will be automatically imported into the repository including the detailed data about unit and service users taken from other documents as DCM spreadsheets store incomplete data (this activity has not been done in this study).

5.4.2 DCM Data Warehouse

A data warehouse [94] is a purpose built data repository to store a historical, consistent and large amount of quality data in a dimensional format for analytical and decision making purposes. The purpose of DCMDW is to provide the users with good quality data, taken from heterogeneous sources, in a consistent format which can be queried to extract the required information for analysis and reporting. The construction of the DCMDW will enable care providers and policy makers to have access to multidimensional data to make decisions to improve the quality of life and standards of care of people with dementia in formal dementia care settings. Researchers and practitioners can also have access to historical, consistent, non-volatile and time variant data to enhance their research and knowledge about the care quality standards. DCMDW will consolidate DCM quality data which will be used for benchmarking in different health care sectors and for researchers.

A dimensional view of data will enable users to look at a large number of inter-dependent aspects of the data from different analytical angles for analysis and decision making activities [102]. Star schema as a dimensional model is recognised as an effective schema for data organisation in most data warehouses [103]. The star schema consists of fact table/s and dimensional tables. A fact table consists of numeric data which contains the data at its atomic level. Dimension tables, linked to the fact table, express the details and justification of the facts for rich and detailed query function. For example the fact about a service user having BCC and ME on a particular date can be found in relation with dimension tables to support this fact. The fact tables can have lowest level of information or aggregated summarized information in them. The fact table can be organised at its lowest granularity level information where BCC and ME of each service user in a specific time/date in a specific unit can be determined. This atomic level of data arrangement in the data warehouse facilitates the data mining and data analysis application on it. As most recent data mining applications support the details at the lowest grain to find the trends and patterns in the data.

The star schema shows the fact table and dimension tables in the DCM domain where dimension tables provide the attribute details which are important to measure the fact in the fact table. For example unit dimension provides the attributes about the unit where mapping has taken place. This dimension can be hierarchically expanded to show the full address of the unit including the area and the country name. This will help to organise the units according to the areas and countries in the querying process.

DCM data from international database will be extracted and integrated with the anonymous data taken from other DCM organisations. Due to the sensitivity and security issues with DCM data, there can be disagreements regarding data provision into DCM international database by the DCM organisations from other countries. Our proposed system solves this problem by allowing these DCM organisations to provide anonymous data into DCMDW for global analysis. This data will be used by different users i.e. governmental authorities or researchers who need large amount of global data on DCM to analyse the quality of care provided in different care homes.

ETL [104] process constitutes different tools to extract the data from local DCM databases and other sources and transform this disparate data into a consistent and compatible format. This data is loaded into DCM data warehouse and used for multidimensional analysis.

5.4.3 Differences between DCM International Database and DCM Data Warehouse

Basically both DCM international database and DCM data warehouse are data repositories which serve different objectives. These objectives are outlined inTable 2:

Table 2: Comparison between International DCM Database and DCM Data Warehouse

Features	**DCM International Database**	**DCM Data warehouse**
Users	Mapper, DBA, Clerk, System Administrator	Managers, Executives, Researchers, Analyst
Function	Basic analytical operations	Complex analytical operations
Data Storage	Current and short term	Historical and long term
Characteristics	Operational processing	Informational processing
Orientation	Day-to-day transaction	Analysis
Access	Read, write, update, delete	Read and update only
Database Design	Application-oriented (ER)	Subject-oriented (star/snowflake schema)
Focus	DCM data in	Information out
Data View	Relational data	Aggregated, summarized data
Data Granularity	Detail	Detail and summary
User Access Methods	Predefined	Predefined and Ad-hoc
Response Time	Fast	Fast to Moderate

5.4.4 Components of DCMDW

When data from transactional databases (in our case it is a DCM international database) or other formats is moved or transformed into data warehouses, it should be arranged into a structure which supports future decision making processes or decision support analysis. The process of developing a data warehouse involves the transfomation of data from Online

Tansactional Processing (OLTP) format to multidimensional fomat to support Online Analytical Processing (OLAP) analysis. Different steps are as follows:

- Data extraction from disparate data sources e.g. OLTP or legacy data sources, flat files, spreadsheets. (This process has been divided into two steps: integration of some data sources into DCM international database and secondly integrating anonymous data taken from other databases, through ETL process, into DCMDW).
- This extracted data is stored into flat files and put together into a staging area (staging area is a place where data extracted from various data sources is kept for performing further actions on it, for example, data scrubbing, aggregating and prepared for loading into data warehouse)
- Loading the data into a data warehouse.

This whole process is also called Extraction, Transformation and Loading (ETL).

5.4.5 Extraction, Transformation and Loading (ETL)

ETL [104] process involves the data extraction from heterogeneous sources, transformation into consistent, compatible and quality data and loading this data into the data warehouse for analysis purposes [105]. It is usually recommended to perform data extraction function from each source system on its own computing format [106]. This action will reduce the data quality issues. The proposed system allows data vendors to input the DCM data into the International DCM database using standard web-based interface or provide anonymous data (already extracted in flat files from their computing platforms) for loading into DCMDW. The extracted data files from international DCM database and other databases will be kept in staging area. Inconsistencies in the data is checked and data is prepared to load into DCMDW.

For practical experiment SQL Server Integration Services (SSIS) was used to extract the data from the DCM international database, transformed it and loaded it into the DCM data warehouse (DCMDW). For this purpose different dimension tables were created and populated using SSIS's Data Transformation Service (DTS). The created dimension tables and fact tables are as follows:

DimBCC, DimDate, DimMapper, DimMappingSession, DimME,
DimServiceUser, DimUnit, FactDCM, FactWIB.

Where DimBCC, DimDate, DimMapper, DimMappingSession, DimME, DimServiceUser and DimUnit are dimensional tables which include the detailes of each of these entities. FactDCM and FactWIB are the fact tables containing numerical data about the facts need to be measured about DCM and WIB (Well/Ill being). During transformation process each dimenson table is given a different id called source id.

During the transformation process a few attribute columns are removed from different tables which are of no use or no longer required for desired analysis process. Data is checked and inconsistencies from the data were removed. The ETL process comprises the fololwing steps:

5.4.6 Data Scrubbing

ETL process in data warehouse allows us to merge the data taken from different sources into a single consistent format. During this process the unseen disparaties and inconsistencies in data can be found. For example misspellings, incomplete information, reused primary keys, nonunique identifiers, input errors and inproper use of address and names. These inconsistencies have to be addressed and solved before loading data into DCMDW. This process of removing inconsistencies from data is called data scrubbing [107].

5.4.7 Aggregating Data

Different analytical tools can be applied on the DCMDW for analysis, reporting and data mining purposes. On Line Analytical Processing (OLAP) [108] tools require summary data or aggregated data rather than detailed data. For example an organisation manager needs to find out about the average well/ill being (WIB) score of all service users of a specific age in all units organised by his organisation from 2000 to 2006 time period. This query can only be solved if he has access to a summary or aggregated data on all service users WIB score from all these units. For this purpose the data is aggregated before loading into DCMDW and organised under different fact tables. These fact tables will be designed according to DCM users requirements. The degree of DCM data aggregation depends on the following three requirements:

- The level of granularity required for DCM analysis from various users
- The speed requirement of international database system

- The type of analysis system there is going to be integrated with the DCMDW (e.g. OLAP, Reporting Services, Data Mining Tools)

5.4.8 Fact Tables

Fact tables provide multidimensional views on data according to the user's requirements [106]. Different fact tables are created to query DCM data at different granularity levels. Granularity represents the level of details in the fact table. The fact tables created at lowest level of details are called base level fact tables . On these fact tables the queries of drill-down and roll-up can be performed efficiently to answer a variety of questions from the data. Fact tables can also be created showing summarized or aggregated levels of details.

5.4.8.1 Fact Tables at Lowest Granularity

Fact DCM (shown in Figure 11) is created at the lowest grain level where fact about each service user having BCC's and ME's, PD's and PE's can be extracted according to the mapping session in a particular unit and in a particular area.

To create the fact table (Fact DCM) a SQL query is written as follows.

```
TRUNCATE TABLE dcmdw.dbo.FactDCM
GO
INSERT INTO dcmdw.dbo.FactDCM
SELECT
  ds.ID as ServiceUserID,
  du.ID as UnitID,
  dm.ID as MapperID,
  dd.ID as DateID,
  dms.ID as MappingSessionID,
  di.TimeFrame,
```

```
  di.BCC,
  di.ME,
  di.PDPE
FROM
  dcm.dbo.DataItem di
  ,dcm.dbo.MappingSession ms
  ,dcm.dbo.Participant pa
  ,dcmdw.dbo.DimDate dd
  ,dcmdw.dbo.DimMapper dm
  ,dcmdw.dbo.DimServiceUser ds
  ,dcmdw.dbo.DimUnit du
  ,dcmdw.dbo.DimMappingSession dms
WHERE
  di.MappingSessionID = ms.ID
  AND di.ParticipantID = pa.ID
  AND dd.Date = ms.SessionDate
  AND dm.SourceID = ms.MapperID
  AND ds.SourceID = pa.ServiceUserID
  AND du.SourceID = ms.UnitID
  AND dms.SourceID = ms.ID
```

The created fact table and dimension tables are shown in Figure 11.

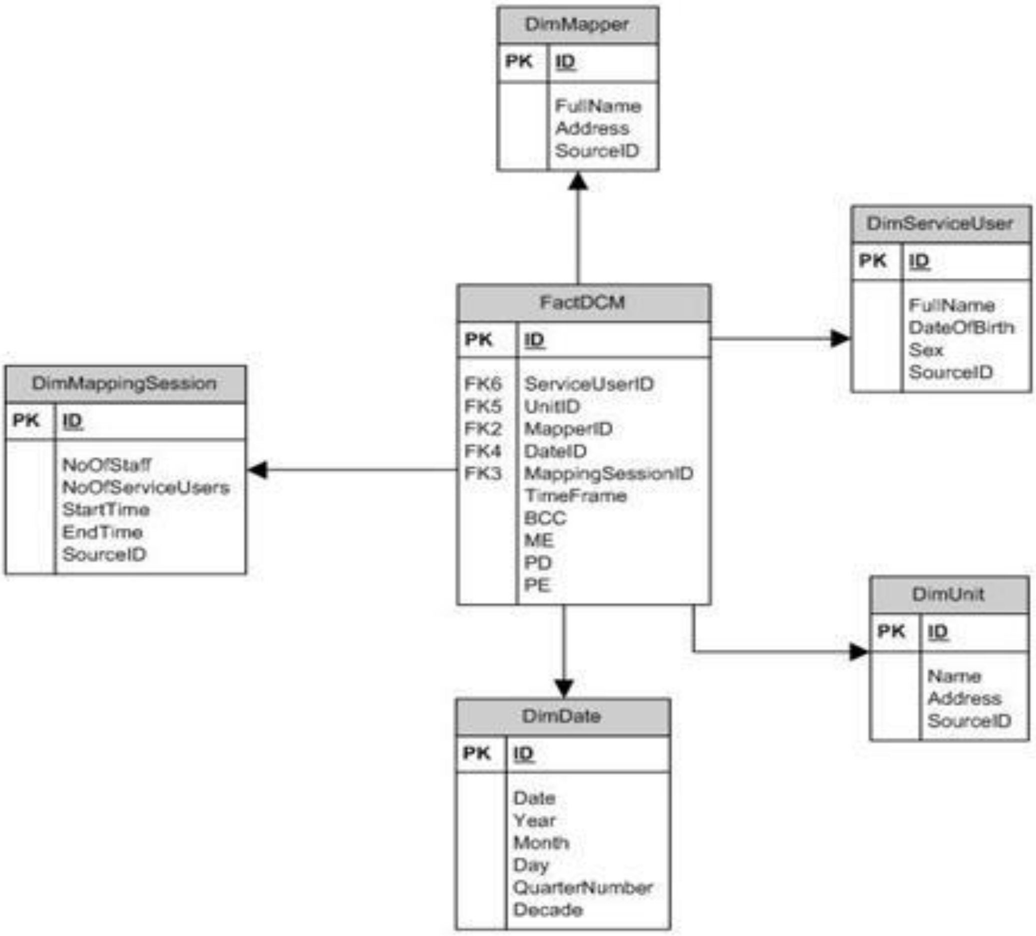


Figure 11: A Star Schema for DCMDW Data Model

Different questions can be answered on this fact table by writing different queries. For example users can find out the most common BCC's recorded in specific years (such as from 2007 to 2009) during various mapping sessions.

Query 1: What are the five most common BCCs for maps conducted during 2007, 2008 and 2009?

The query is as follows:

```
SELECT TOP 5
  BCC,
  COUNT(*) as BCCCount
```

```
FROM    FactDCM JOIN DimDate ON FactDCM.DateID = DimDate.ID

WHERE Year IN (2007,2008,2009)

        GROUP BY BCC

        ORDER BY BCCCount DESC
```

The five most comon BCC are B,N, NULL (represents unmapped), F, A as shown by the query results below:

	BCC	BCCCount
1	B	1547
2	N	1241
3	NULL	1044
4	F	744
5	A	579

Users interested to know the percentage of +3 ME's recorded about all service users in a group level in all mapped units in the given data set in particular dates can extract the required information from the designed fact table by applying the following query:

Query 2: How many mapping sessions were done in UK Trust Homes in 1999 & 2000?

The query is as follows:

```
SELECT
  dd.Year,
  COUNT(DISTINCT MappingSessionID) as [Number of mapping sessions]
FROM
  FactDCM fd,
  DimDate dd,
  DimUnit du
WHERE
  fd.DateID = dd.ID AND
```

```
        fd.UnitID = du.ID AND
        du.IsTrustHome = 1 AND
        dd.Year IN (1999,2000)
    GROUP BY
        dd.Year
```

Results of this query are:

	Year	Number of mapping sessions
1	1999	9
2	2000	6

Query 3: What is the % of +3's MEs scored on a group level across all units in maps from 1 Jan 2008 - 30th June 2008?

```
SELECT
        'Unit ' + convert(varchar,MAX(du.ID)) as [Unit],
        MappingSessionID,
        100* CAST(COUNT(CASE ME WHEN 3 THEN 1 ELSE NULL END) AS
        float) / COUNT(ME) as [Percentage of +3's]
FROM
        FactDCM fd,
        DimUnit du,
        DimDate dd
WHERE
        fd.UnitID = du.ID AND
        fd.DateID = dd.ID AND
        (dd.Date >= ‘1 JAN 2008’ AND dd.Date <= ’30 JUN 2008’)
GROUP BY
                fd.UnitID, MappingSessionID
```

The results are as follows:

	Unit	MappingSessionID	Percentage of +3's
1	Unit 8	13	0
2	Unit 9	14	1.67597765363128
3	Unit 9	15	11.1111111111111
4	Unit 9	16	4.89296636085627
5	Unit 11	17	0
6	Unit 11	18	15.8054711246201
7	Unit 9	19	9.02527075812274
8	Unit 11	20	4.23076923076923
9	Unit 11	21	2.43055555555556
10	Unit 9	22	8.01186943620178
11	Unit 9	23	2.8125
12	Unit 8	24	0.56980056980057
13	Unit 9	25	0.308641975308...
14	Unit 10	26	5.60747663551402
15	Unit 10	27	9.94152046783626
16	Unit 12	28	0
17	Unit 13	29	4.08921933085502
18	Unit 14	30	21.455938697318
19	Unit 15	31	24.7422680412371
20	Unit 16	32	2.6431718061674
21	Unit 17	33	20.2290076335878

This query is performing an aggregation on the FactDCM table by Unit and Mapping Session. This aggregation can be performed during the the ETL stage and stored as a new fact table to simplify the query even further. However for this particular question doing aggregation inside the query offerred the most flexibility.

The process of data anonymization can be a part of ETL process where we can identify different levels of anonymization. Given the DCM fact table (Figure 11) includes the anonymous data about all service users (identifiers are used) where queries can be applied on anonymous data. In dimension tables we have details of the fact described in the fact table. The full details of service users, mappers and units are given in only dimension tables, which can only be extracted if user wants to have a detailed view of data. Different users can be authorized to have access to only those fact tables which have anonymized data about service users, mappers and units.

There are two main advantages of creating fact tables at the lowest granular level in the context of the DCMDW. One is that users can extract data at its very basic level and the other is the application of data mining tools on these tables. The drawback of this is that the fact tables at lowest granularity level require a large space for data storage, as lowest grain means a large number of fact table rows [106].

5.4.8.2 Fact Tables at Highest Granularity

Fact tables can be created showing aggregated facts to support queries based on summary numbers [109]. A measure called the well or ill being (WIB) score is an important aggregated calculation required for the DCM analysis [11]. WIB is the calculated average measure of all ME's of an individual service users or a group of service users during one mapping session, The WIB score can describe the overall well being or ill being levels of the person with dementia in dementia care settings. A fact table FactWIB is created to calculate this average measure in service users. The query is shown below:

```
TRUNCATE TABLE dcmdw.dbo.FactWIB

GO

INSERT INTO dcmdw.dbo.FactWIB

SELECT

  ds.ID as ServiceUserID,
  dms.ID as MappingSessionID,
  du.ID as UnitID,
```

```
  dm.ID as MapperID,
  dd.ID as DateID,
  di.WIBScore as WIBScore
FROM
  ( SELECT ParticipantID,  MappingSessionID,
  CAST(Round(AVG(cast(ME as decimal)),1) as decimal(19,1)) as WIBScore
  FROM dcm.dbo.DataItem
  GROUP BY ParticipantID, MappingSessionID) as di
  ,dcm.dbo.MappingSession ms
  ,dcm.dbo.Participant pa
  ,dcmdw.dbo.DimDate dd
  ,dcmdw.dbo.DimMapper dm
  ,dcmdw.dbo.DimServiceUser ds
  ,dcmdw.dbo.DimUnit du
  ,dcmdw.dbo.DimMappingSession dms
WHERE
  di.MappingSessionID = ms.ID
  AND di.WIBScore IS NOT NULL
  AND di.ParticipantID = pa.ID
  AND dd.Date = ms.SessionDate
  AND dm.SourceID = ms.MapperID
  AND ds.SourceID = pa.ServiceUserID
  AND du.SourceID = ms.UnitID
  AND dms.SourceID = ms.ID
```

The FactWIB (fact table) and its dimension tables are shown in Figure 12.

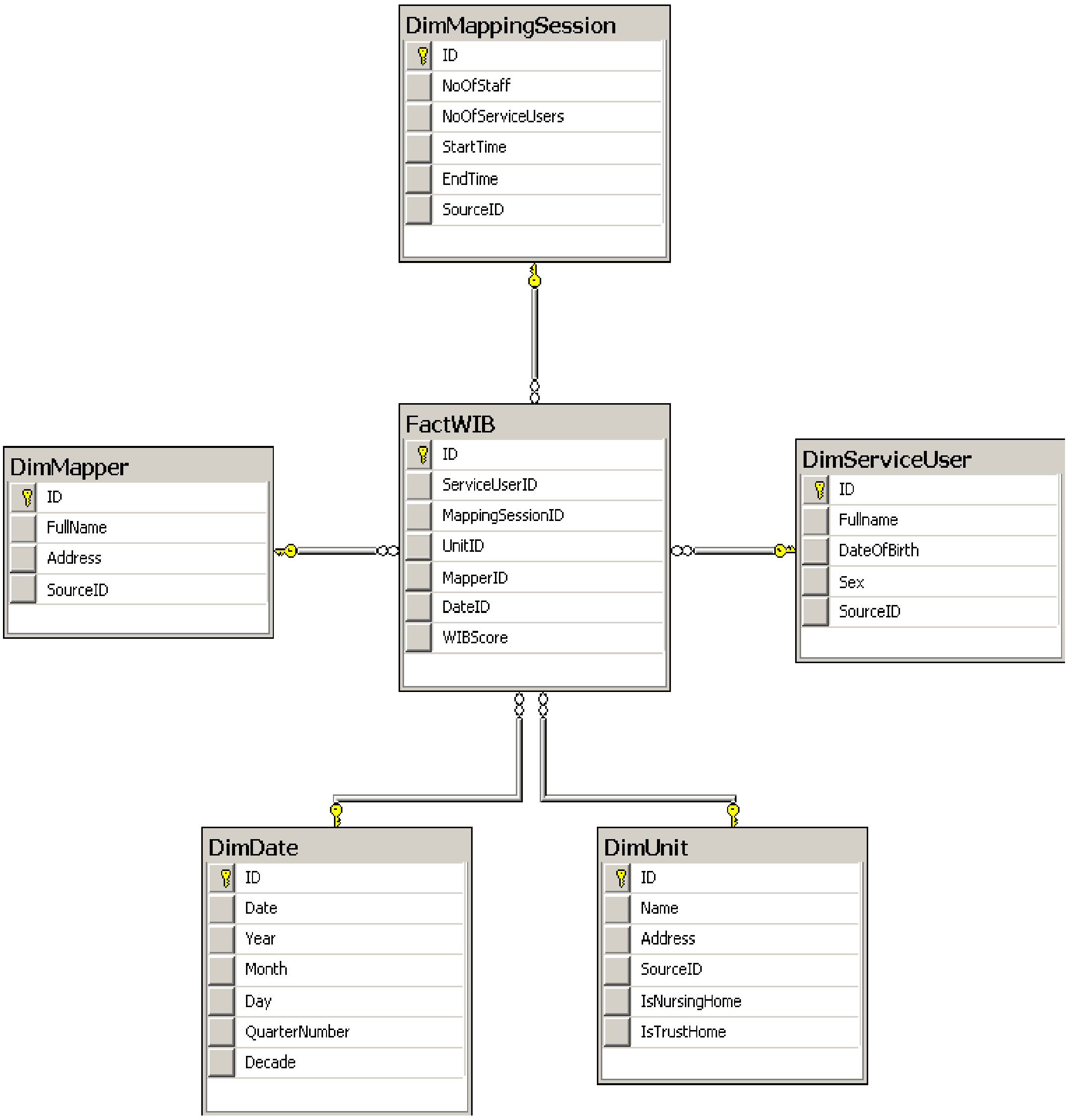


Figure 12: A Star Schema Showing Fact WIB Table and Dimension Tables

Different querries were run to answer questions which could not have been answered before using existing DCM data management system (spreadsheets).

For example,

Query 4: What is the average group WIB score for maps conducted during 1999?

The query to answer this question is as follows:

```
SELECT
  AVG(fw.WIBScore)
FROM
  FactWIB fw,
  DimDate dd,
  DimUnit du
WHERE
  fw.DateID = dd.ID AND
  fw.UnitID = du.ID AND
  du.IsNursingHome = 1 AND
  dd.Year = 1999
```

The average group WIB score is 1.1 in the year 1999 as show by the results below:

	AverageWIB
1	1.106896

Query 5: What are average WIB Scores of all female service users in Trust Homes and Care homes ?

The query is as follows:

```
SELECT
  'Nursing Homes' as Unit,
  AVG(fw.WIBScore) as Score
FROM
```

```
  FactWIB fw,
  DimUnit du,
  DimServiceUser su
WHERE
  fw.UnitID = du.ID AND
  fw.ServiceUserID = su.ID AND
  su.Sex= 'F' AND
  du.IsNursingHome = 1
UNION
SELECT
  'Trust Homes' as Unit,
  AVG(fw.WIBScore) as Score
FROM
  FactWIB fw,
  DimUnit du,
  DimServiceUser su
WHERE
  fw.UnitID = du.ID AND
  fw.ServiceUserID = su.ID AND
  su.Sex= 'F' AND
  du.IsTrustHome = 1
```

Results are shown below where Nursing Homes are having average WIB score 1.17 and Trust Homes having 0.578.

	Unit	AveWIBScore
1	Nursing Homes	1.171428
2	Trust Homes	0.578000

The above query could be modified to find out the Nursing Home and Trust Home’s locations by joining on the DimUnit dimension table at the address attribute.

Different fact tables and dimension tables are created to answer the variety of user's queries from DCM data. An architecture of multidimensional DCM data showing some developed fact and dimension tables is shown in Figure 13.

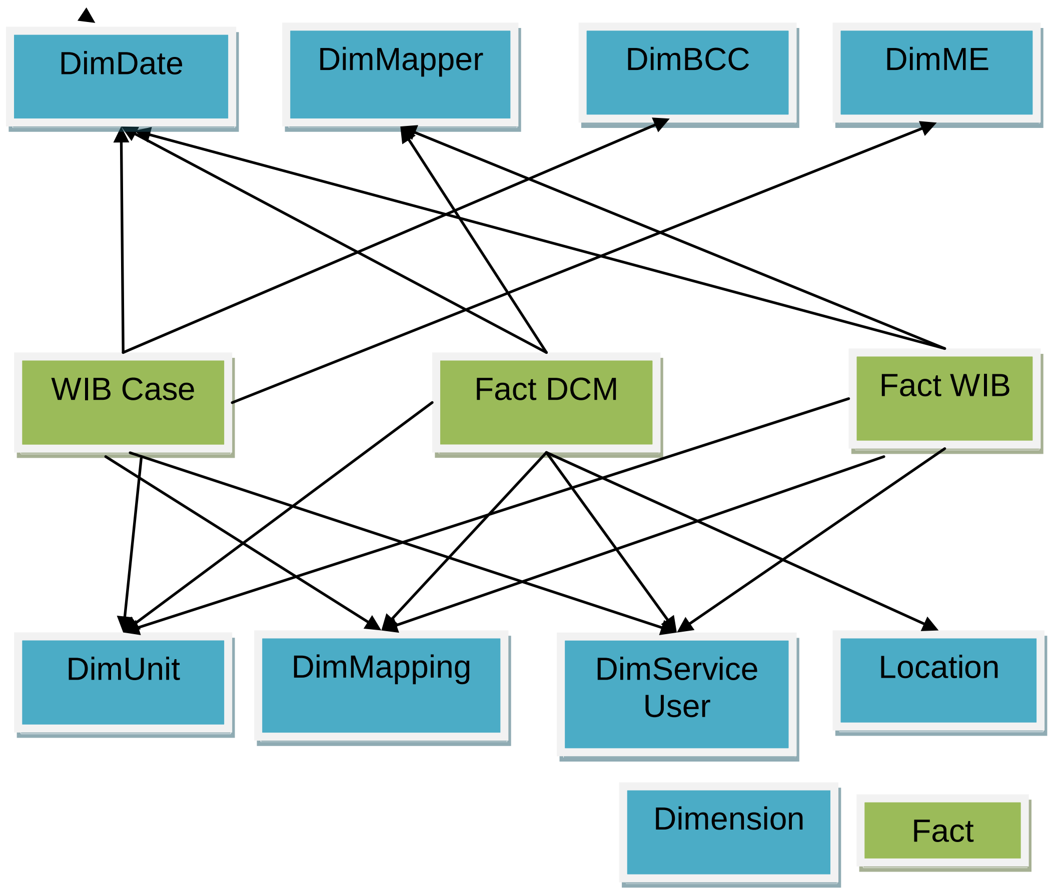


Figure 13: Multidimensional Architecture of the DCM Data in DCMDW

In this figure different fact tables are showing their relationship with the dimension tables. For example fact table (Fact WIB) is showing association only with the tables containing data about Date, Unit, Mapping session, mapper and Service users. There is no association with the BCC and ME dimension tables. The reason is that the table FactWIB table contains the aggregated measures of BCC and ME already calculated in WIB Score attribute (taken BCC and ME values from Data Item table). This already calculated aggregated measure on WIB Score reduces the time of calculation.

5.4.8.3 Applications on DCMDW

Data warehouse provides an ideal environment for multidimensional analysis [106]. Different analytical tools i.e. Online Analytical Processing (OLAP) [106], Data Mining (DM) and Decision Support Systems (DSS) [103] could be equipped with the data warehouse to provide different users with multidimensional analysis, patterns recognition and decision making [102]. The OLAP tools provide a user with the facility to drill down to the atomic level or roll up to the aggregated data for summaries [110]. A user is able to analyse the data from different angles. For example a DCM user will be able to find out information such as the WIB score of each service user in a particular unit in a specific time period, or in a particular year the details of all mapping session undertaken by a particular mapper etc. DCMDW will enable users at different authority levels to customise the process of information retrieval using multidimensional view of data on OLAP tools.

OLAP applications are used to interrogate the DW or decision support systems to find out the answers. Different analytical systems are used for this, for example Online Analytical Processing (OLAP), Multidimensional Online Analytical Processing (MOLAP), Relational Online Analytical Processing (ROLAP) [111]. The OLAP and MOLAP tools are used on the data which is aggregated in a separate database for querying while ROLAP is used to query the relational database directly.

Data can be arranged multidimensionally to provide a view from different angles. DCM users can see the DCM data in a particular time period, collected from different types of dementia units internationally. OLAP tools also help users to drill down data to the lowest granular level. As shown in Figure 14 DCM data is drilled down on address attribute where mapping data acquired from different parts of the UK's dementia care settings in different time periods can be viewed.

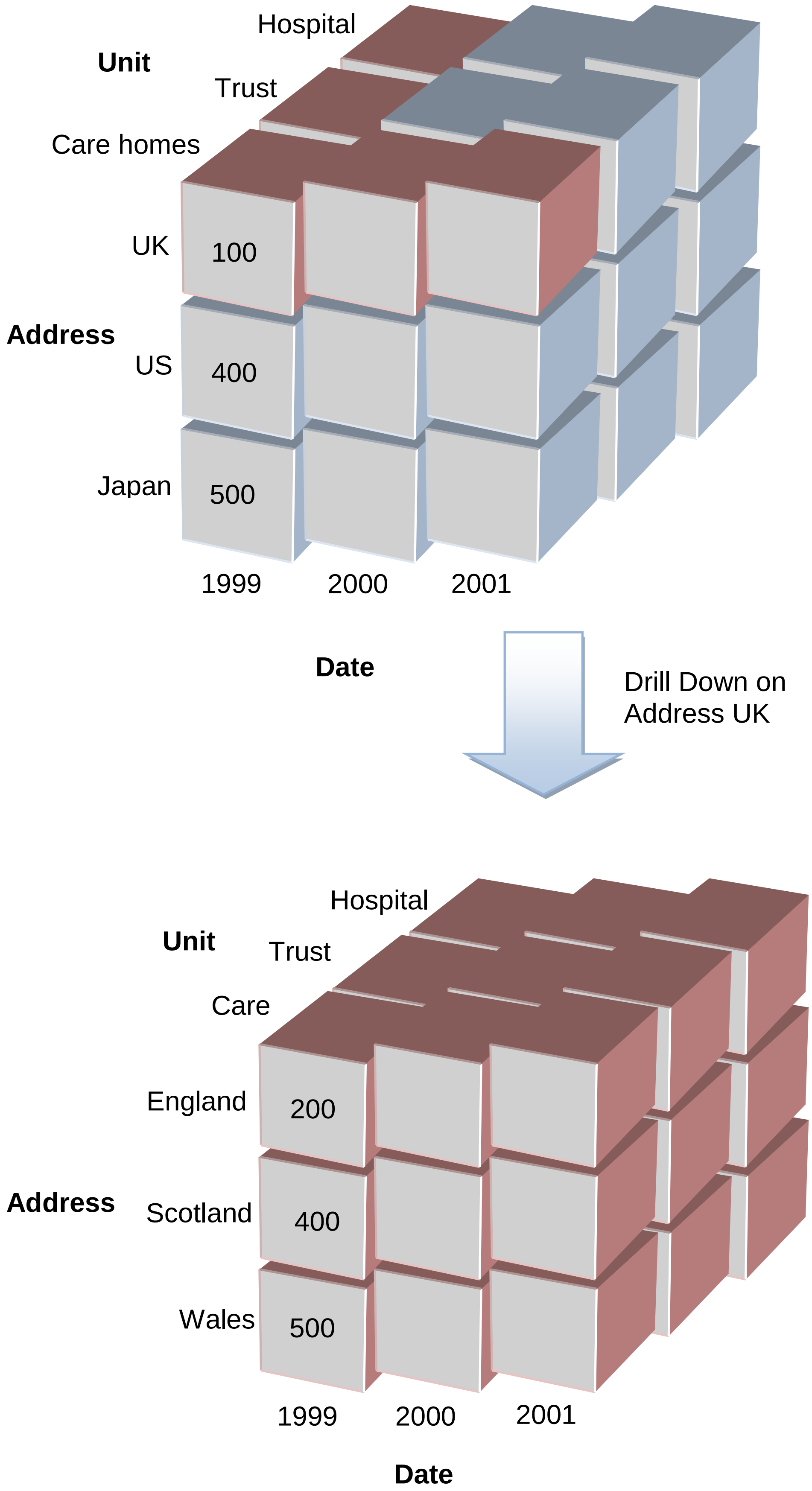


Figure 14: DCM Data Drill Down at Address Attribute

Data can be drilled down further to see the mapping sessions which took place in each dementia unit type in England. In the same manner data in multidimensional cubes can be rolled-up to see the aggregated summary of the data.

5.5 Data Mining and Data Warehouse

Discovering important information from large databases is becoming an important and essential part of business requirements. Data Mining (DM) is a process of extracting the valuable knowledge from a large amount of data.

Data warehouse is used by business executives and analysts to perform data analysis and make strategic decisions. Data warehouses are used as a Plan-Execute-Assess “Closed -loop” system to manage any organizations data management process [118]. The basic function of DW is to store historical, summarized data for analysis and report generation but this functionality is progressing to provide clean and structured data for strategic purposes, multidimensional analysis and data mining.

For data mining on any data, the subject data needs to be clean, of good quality and available in granular form so patterns can be extracted at any levels of data granularity from different aspects. A data warehouse provides an integrated environment with historical data stored at different granularity level and abstraction levels for data mining techniques.

DCM data is a combination of a variety of data. Demographic data of people with dementia and mappers is collected alongside the codes which represent the behaviour and mood/engagement of service users who participated in the mapping session. The DCM coding data is complex and rule-based and cannot be used for mining purposes unless pre-processed to extract the basic meanings of attributes. For example to find out the positive engagement and distress level in mapped participants, correct BCC's should be associated with correct ME's. This topic is discussed in detail in Chapter 6.

5.5.1 Conclusions

In this chapter the DCM data management framework based on a data warehousing approach is proposed, which accomplishes a major component of the DCM data governance framework. The described data management framework deals with the data capturing, storing and dissemination issues in the DCM system and provides the solution to capture, store and analyse data from national and international DCM organizations. Different steps in designing the management framework (DCM International Database, DCMDW) have been described and evaluated by experimental work on a small scale level with a limited data provided. I developed a web application as a front end to the designed and developed DCM international database. Through this application DCM data was captured and stored in a database for basic analysis. This data was loaded into the DCM data warehouse through a ETL process. Different queries are applied to show the data retrieval from the DCMDW at the lowest and highest levels of granularity. Results are encouraging. We provide the basic structure of fact and dimension tables on the DCM data; however more fact tables, representing different required data views, can be applied to the DCM data for complex analytical questions as well.

Chapter 6: Application of Data Mining on Data Warehouse

6.1 Introduction

This chapter presents data warehousing and data mining connectivity and explains how the data from DCMDW can be used for data mining purposes. A case study, describing the data mining task (clustering) used on DCM data to identify the groups or classes in DCM data for further analysis, is also explained.

6.2 Data Warehousing and Data Mining

To accommodate the facility of extracting important information from the data for a variety of users, users should have an opportunity to mine or extract the data at different abstraction levels [112]. DW provides facilities to users to extract required data or information from pre-arranged data at different levels of granularity [106]. Extraction of important information from a variety of databases and other data sources is a complex and time consuming process in data mining applications. A data warehouse facilitates this by giving an integrated data collection environment. Although data mining techniques can operate on any kind of unprocessed or even unstructured data, the DW environment facilitates the DM applications to be applied efficiently and effectively on the data arranged at different granularity levels. It provides more in-depth and often more multidimensional knowledge [113].

DM can be considered to be a threat to privacy as patterns can be found about a person by taking data from different data sources [112]. A data warehouse containing only anonymous data can reduce or eliminate this issue and DM applications can have access to data which is permissible for further analysis and pattern recognition. Different approaches and steps [114] are indentified by different researchers for DM processes. All these suggestions are based on more or less same basic steps. For example: pre- processing data (collection, sampling, cleaning, modelling), applying different DM algorithms on data to find un- identified patterns in the data. Data is collected from different data sources and then through the ETL process is cleansed, transformed and loaded into DW for further multidimensional analysis and DM applications.

6.3 Data Mining Techniques

Data warehouses provide a convenient environment for performing multidimensional data analysis and data mining tasks [115]. Different data mining techniques can be applied on DCM data to extract important information and find un-identified patterns in data. Some are described below:

6.3.1 Association Rules

This method reveals association links between different attributes/values in one transaction within databases [116]. Association rules can be applied on DCM data to associate the service user’s well or ill being with the weather condition of that day when mapping took place.

6.3.2 Clustering

Clustering [117] is an un-supervised learning technique in data mining to classify data into groups (clusters) based on the similarity in objects. Similar objects are grouped in one cluster. These objects are dissimilar to other objects grouped in different clusters. Clustering is applied on data to get in-sight into large amount of unclassified or meaningless data to make sense out of it. This technique is used to pre-process the data as well for further analysis.

6.3.3 Classification

Classification is a supervised learning technique in data mining which is applied to labelled categorical data. In this method the training sets/samples with output classes are used to classify new data. There are different techniques for data classification such as decision tree classification, Bayesian classification, rule-based classification, neural networks and support vector machines [118].

6.4 Clustering

Clustering is a process to group a set of physical or abstract objects into classes of similar objects [118]. A cluster presents a collection of data objects which are similar to each other but at the same time are dissimilar to other data objects in other clusters. Multivariate techniques such as cluster analysis may allow researchers to identify groups, or clusters, of related variables. Large data sets with multiple dimensions can be reduced to a few clusters or groups for subsequent analysis.

Cluster analysis is based on probability measurements of the variables within the cluster where data is viewed as coming from a mixture of probability distributions [119].

To find the hidden variables that classify our data, cluster analysis is a useful technique in data mining [120]. Cluster analysis is an exploratory data analysis tool. The aim of this tool is to sort different objects into groups where the degree of association between two objects is maximal if they belong to the same group and minimal otherwise. Cluster analysis techniques classify the groups of similar objects without explaining why they exist [121]. This technique is mostly used when we do not have any prior hypothesis or output class to train our data set on it. This is the reason the clustering technique is based on unsupervised machine learning technique. DCM data lacks the output classes and due to the huge amount of data, domain experts cannot classify data into useful groups. Clustering will help us to identify classes or groups in the data according to the similarity of the objects.

6.4.1 Clustering in Medical Data

Clustering techniques have been applied to a wide variety of medical domains [122,123]. For example clustering techniques, on grouping different types of diseases into similar groups, based on the type and severity of disease, cure of diseases, and symptoms of diseases, have been applied effectively in medicine and health [121]. Different diseases such as paranoia, schizophrenia, etc., can be correctly diagnosed based on the clusters made from their important symptoms. In the field of psychiatry clustering analysis is considered to be beneficial for a successful therapy [121]. Medical data is considered to be complex data stored in large databases. Different types of clustering methods can be applied on medical data to classify it into meaningful groups for further analysis. Ensemble clustering technique has been introduced in [124], where the authors discuss the effectiveness of this approach in medical diagnostics. Partition method and hierarchical methods are also used in clustering medical data into groups of similar objects. Whenever we need to classify a large amount of un-classified data, a clustering technique for analysis is considered to be of great benefit.

Some other benefits of apply clustering techniques on the data are as follows:

6.4.1.1 Grouping Similar Data using Clustering

There are different usages of clustering – one is the grouping of data according to the similarities in variables. For example cluster analysis can be used to learn more about the type of service users in any care setting based on their similar behaviours, and even the type of care settings can be grouped based on their service user's well/ill being and the quality of care provided in that care setting.

6.4.1.2 Anomaly Detection using Clustering

Another usage is to find anomalies in the data using clustering technique. When issuing a query to the cluster model, the model determines the cases belonging to specific clusters and those cases which do not belong to any cluster may be alarming. Leveraging the properties of the clustering algorithm, we can determine the data cases which do not fit any model and represent an anomaly or bad data. Figure 15 shows the data anomaly cases in the data.

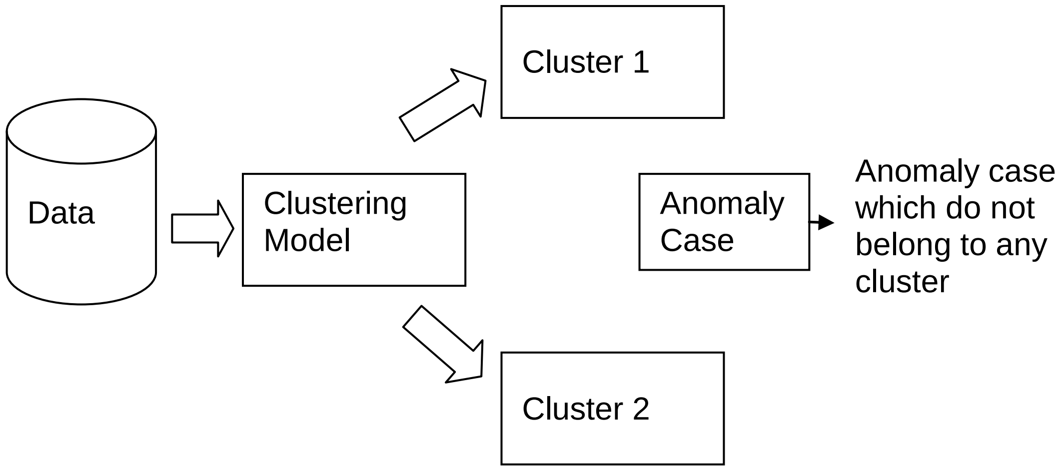


Figure 15: Clustering: To Detect Anomaly Cases in Data

6.4.1.3 Clustering as an Analytical Step

By applying cluster analysis, we have data in groups representing similar characteristics in the variables. By grouping data, we can create better models to answer the analytical questions in depth. For example when we group data according to a service user's behaviours in particular care setting, it is easy to concentrate on that particular group to answer the analytical questions based on the recognised behaviour type and care setting type and we can tightly focus on the reason why this type of behaviour is common in the particular care setting and why a particular gender is behaving in a particular way. Perhaps we can create a tree model to predict which gender of which age group will more likely be able to show positive behaviour in which care setting or can be happy in which care setting, (See Figure 16).

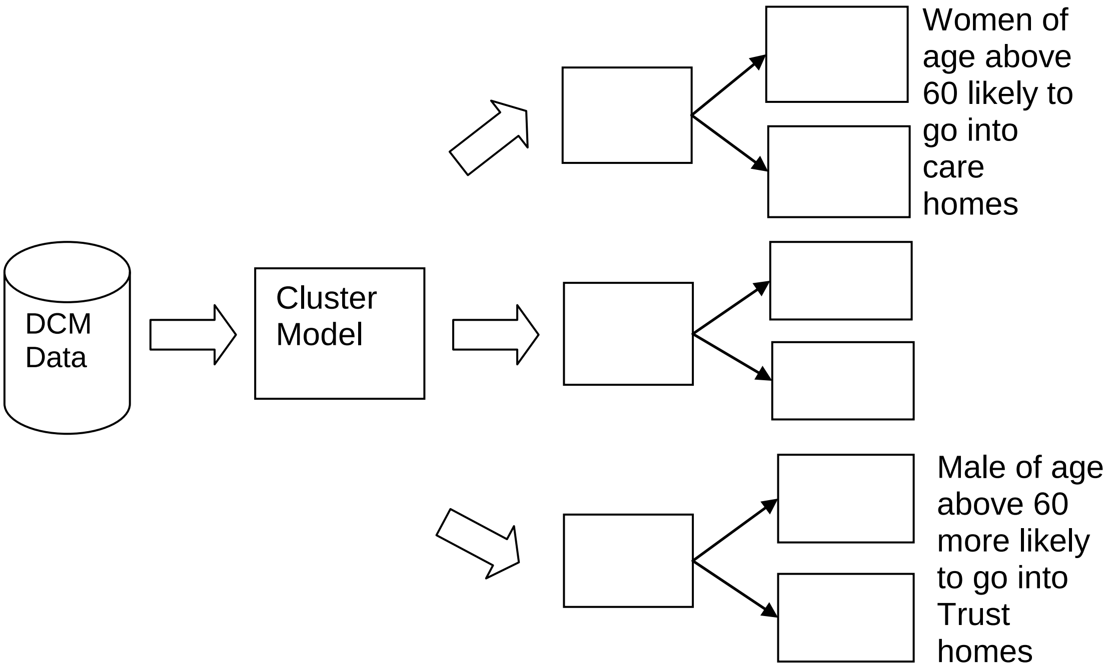


Figure 16: Clustering: A Pre-processing Step for Decision Trees

6.5 Different Types of Clustering Methods

Cluster analysis can be done using different methods. The choice of a particular method depends on the type of data used for analysis and the question we need to answer from the data. Different popular clustering methods are briefly described next.

6.5.1 Partitioning Methods

Given a data set of ‘n’ objects, a partitioning method divide a data set into ‘k’ parts where each part represent a cluster containing similar variables and where k<= n.

This partition is based on the fact that each group must contain at least one object and each object must be a part of some group. This method generates an iterative process of partitioning of the data set until all the similar objects or closely related objects are placed in one group. The K-means is a popular algorithm used under this clustering method. In k-means clustering, an object belongs to a cluster based on the centre point it is closest to. The distance of each object from the centre of the cluster is measured using Euclidean distance. When all objects have been grouped into clusters the centre of the cluster is shifted to the mean of all assigned objects. K-means clustering is also considered as “Hard Clustering” as each object is assigned to one and exactly one cluster. Expectation Maximization (EM) [127] is a “Soft Clustering” method where data is partitioned based on their attributes’ probability distribution. This method does not place exactly one object in one cluster but similar objects are placed in one cluster based on their highest probability.

6.5.2 Hierarchical Methods

In this method the data set is divided and subdivided or partitioned in hierarchical order. A hierarchical method can further be divided into agglomerative as bottom-up approach or divisive as top-down approach based on the hierarchical decomposition approach. Bottom-up approaches start with each object forming a cluster and then gradually merging into other clusters based on the similarity or closeness of other objects. Top-down approaches start with all of the objects in the same cluster and then splitting up into smaller clusters until each object is placed into a single cluster.

6.5.3 Density-based Methods

The density-based clustering method is devised to handle a data set that includes or generates arbitrary clusters. The idea is to continue growing the given cluster until all of its data objects (density) exceeds a threshold. This method is good for checking the data set and recognising clusters of arbitrary shapes. DBSCAN [125] and OPTICS [126] are popular techniques used in this kind of method.

6.6 A Case Study

In any dementia care setting the assessment of care environment for the potential it creates for positive engagement by service users is a valuable indicator to assess the quality of care provided in that care setting. In DCM, positive engagement is seen as one of the key things to improve the quality of life in dementia [12]. In the same manner the level of agitation and distress in service users living in any dementia setting can be assessed from collected DCM data. Different other types of behaviour in people with dementia (service users) can be assessed by processing, analysing and mining the collected DCM data, these other behaviours are being withdrawn and being engaged passively.

A case study is designed to explore the DCM data taken from different dementia care settings on service users, mapped in different mapping sessions in particular care settings, to assess their particular behaviour based on the type of care setting they are living in and their gender. For this purpose we apply clustering on the DCM data extracted from the DCM data warehouse.

The purpose of this case study is to perform exploratory data analysis on the DCM data, and see if the clusters that are generated are meaningful.

6.7 The Data Set

For the case study the experimental data set is obtained from the proposed and developed DCM data warehouse. This data set contains mapping data taken from two dementia care settings i.e. Nursing Home and Trust Home, where 52 service users were mapped during different mapping sessions in different time periods and dates. The number of service users from the Nursing Home was 12 and 40 from the Trust Home. Out of 52 total service users, there were 36 female and 16 male participants involved in mapping. The following attributes/variables, which could answer the required questions, were extracted from the already existing tables stored in DCMDW. These attributes are:

- Service user ID (as object identifier not used in clustering process)
- Service user's gender
- Unit Type
- Category

6.8 Data Pre-processing

DCM data stored in data warehouse is used for clustering. The steps of this data processing are shown in Figure 17.

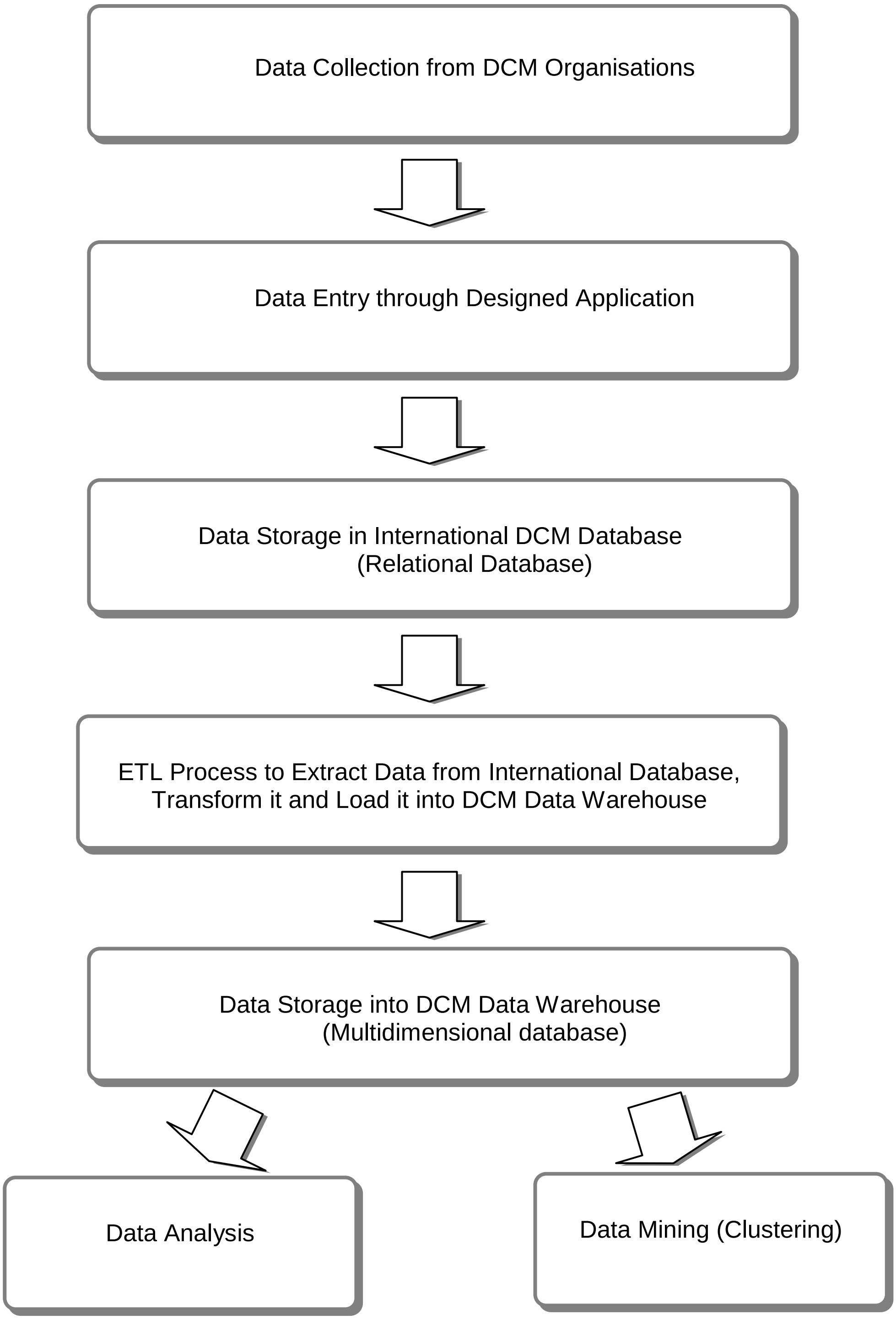


Figure 17: DCM Data Processing Steps

Data is in the pre-processed form in the data warehouse as the data is already complete and free of anomalies. Still further steps are taken to extract the required attributes for clustering. These attributes are identified from tables, stored in the DCMDW. These tables are shown in Figure 18.

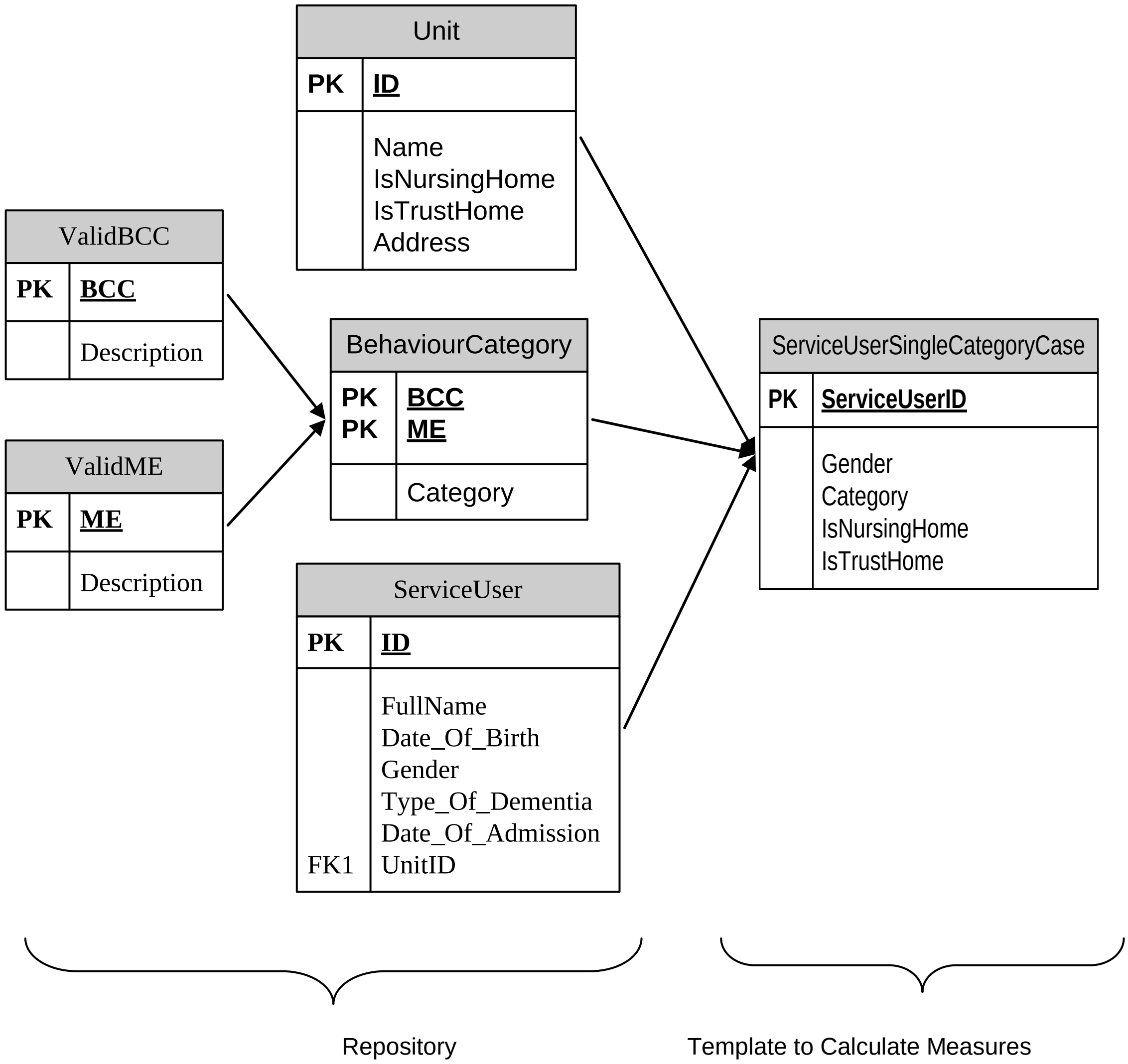


Figure 18: Data Model Showing Tables and Attributes Used in Clustering

ServiceUserSingleCategoryCase table is a template of calculated measures of BCC's and ME's. The BCC's related with specific ME's, together are categorized according to DCM's defined rules. Each service user will show a percentage of time he/she spent in that particular category. Instead of performing clustering on the numerical values of percentages, we calculated the highest percentage of time spent by each service user in each category and extracted the gender and the relevant category. After performing this action we performed the clustering task on all discrete attributes.

The defined DCM categories based on BCC's and relevant ME's according to DCM rules are described as follows:

- Positive engagement (service user is showing positive behaviour and well being)
- Agitation and distress (service user is showing negative and distressed behaviour)
- Withdrawn behaviour (service user being totally uninvolved and disengaged from his/her environment)
- Passive engagement (service users is not actively engaged in any conversation or activity)

6.8.1 Positive Engagement

BCC codes like A, D, E, F, G, I, J, K, L, O R, S, T, V and Y are high potential category codes. These codes can show the potential for positive engagement in an individual or group level in any care setting environment. The percentage of total time spent in each of this high potential category code is calculated to see the overall positive engagement of an individual or a group.

6.8.2 Agitation and Distress

The level of agitation and distress in service users can be assessed by calculating overall percentage of time spent in some specific BCC codes associated with negative ME values. These BCC codes associated with negative ME (-1, -3, -5) are K, W, Y and U. Where BCC code "U" can be associated to positive or negative ME value.

6.8.3 Withdrawn Behaviour

Withdrawn behaviour in service users is represented by the percentage of total amount of time spent in categories C and N collectively.

6.8.4 Passive Engagement

The time spent in BCC "B" is a category of moderate potential. The service user who spent most of their time in this category has the potential for active engagement, but also potential for disengagement. By recognising those individuals who spent most of their time in this category can be picked out and care unit staff can make arrangements to turn this passive engagement into positive engagement.

Below is the Table 3 created by extracting the required attributes from the three tables DimServiceUser, DimUnit, DimBehaviourCategory

Table 3: Service User Single Category Case Table

ServiceUserSingleCategorye
Service user ID
Gender
Is Nursing Home
Is Trust Home

The query to create this table is shown below:

```
Select category, sex , Is TrustHome into ServiceuserSingleCategoryCase Table
select category.*, Sex, IsTrustHome
into ServiceUserSingleCategoryCase
from (
select ServiceUserID, UnitID, Category
  from
          FactDCM fdcm
          join DimME on fdcm.ME = DimME.Code
          join DimBehaviourCategory cat on fdcm.BCC = cat.BCC and DimME.Mood =
cat.MEMood
group by ServiceUserID, UnitID, Category
having COUNT(*) = (
          select MAX(score) as maxscore from (
          select ServiceUserID, Category, COUNT(*) as score
          from FactDCM join DimBCC on BCC = DimBCC.Code
          where ServiceUserID = fdcm.ServiceUserID
          group by ServiceUserID, Category) as innerq
          group by ServiceUserID
)) as category, DimServiceUser, DimUnit
where category.ServiceUserID  = DimServiceUser.ID
and category.UnitID = DimUnit.ID
```

6.9 Application of Clustering Algorithm

The attributes having discrete categorical data type are used for clustering. EM method is applied which is an extension of k-means method and favourably accommodates discrete categorical variables. In this method at first, different probabilities (weights) are assigned randomly to each object for each cluster. In successive iterations, these probabilities are refined (adjusted) to maximize the likelihood of the data given the specified number of clusters. On the other hand the k-means method produces exactly k different clusters of greatest possible distinction. The results of both clustering methods are different. The k-means assigns objects to clusters to maximize the distances between clusters while EM method assigns the objects into specific clusters based on probability distribution. EM clustering method has been applied on the DCM data set to classify the data into classes. There are two reasons to choose this method on the DCM data; one is the suitability of the EM method with discrete and categorical data type and second is that this method represents the soft clustering approach on data which maximize the likelihood of the model [127]. This approach will enable users to see the number of actual population of objects in the data set and the number of objects placed in each cluster based on their probability distributions.

The attributes acquired for clustering analysis are shown in Figure 19.

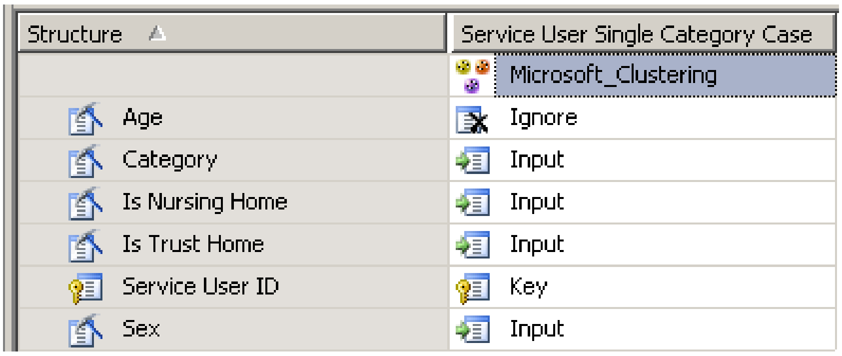


Figure 19: Attributes Used for Clustering

Where Service User ID is used as an object identifier and does not perform any role in the clustering process.

Data is grouped into 5 clusters. After analysing each cluster, it is labelled with a suitable name which represents the related objects in that group. (see Figure 20).

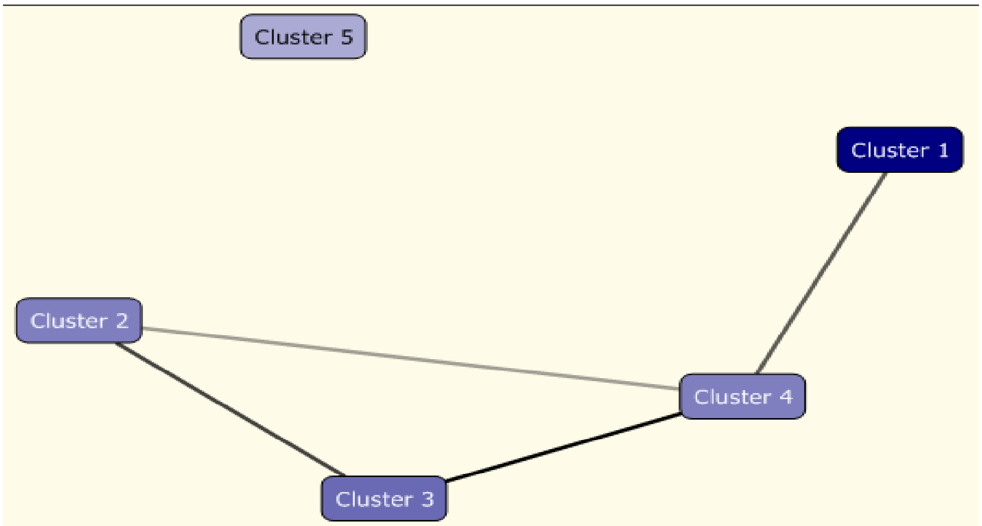


Figure 20: Cluster Model Showing All Links

The clusters are labelled according to the type of data they are divided in. These labels are shown in Table 4.

Table 4: Clusters and their Interpreted Labels

Clusters	Labels
Cluster 1	Females in Trust Home with withdrawn behaviour
Cluster 2	Females in Trust Home with positive engagement
Cluster 3	Females in Trust home with passive engagement
Cluster 4	Males in Trust Home with withdrawn behaviour
Cluster 5	Females in Nursing Home with passive engagement and withdrawn behaviour

In Figure 20 each cluster is displayed as a single node. These nodes are scattered across the field and are grouped automatically based on the similarities in objects. The relative strength of the similarity is shown with a dark and light shaded line. By looking at the clusters we can ask different questions, for example, in which cluster are female service users living in Trust Home or Nursing Home and showing a particular behaviour being placed? We can see the clusters showing week and strong relationships and also those showing no relationship with any other cluster. Cluster 1, 2, 3, and 4 are showing relationship with strongest similarity based on the unit type attribute as they all are living in Trust Homes .The cluster 5 has no relationship with other clusters as this is the only cluster that contains female service users from the Nursing Home. The details of each cluster can be drilled down to see the actual data with attributes and its values (see Figure 21 showing detailed attribute and values of Nursing Home withdrawn females (cluster 5)).

Age	Category	Is Nursing Home	Is Trust Home	Service User ID	Sex
54	Passive Engag...	True	False	79	F
61	Withdrawal	True	False	82	F
56	Passive Engag...	True	False	83	M
67	Passive Engag...	True	False	86	M
65	Withdrawal	True	False	89	F
66	Withdrawal	True	False	98	F
54	Positive Engag...	True	False	100	F
49	Withdrawal	True	False	101	M
54	Passive Engag...	True	False	103	F
67	Withdrawal	True	False	105	F

Figure 21: A Drill through Data of Cluster 5

The cluster categorization can be seen in cluster profile view in Figure 22. The attributes category, isNursing home, isTrust home and sex are presented as rows along with the columns showing each cluster. The attribute sex and its distribution across all clusters can be seen. It is visible that the cluster presenting all population shows the data distribution where service users with withdrawn behaviour and service users of female gender have highest probabilities than male service users.

The above figure shows that the majority of females users having withdrawn behaviour and living in Trust Homes are placed in cluster 1. While the females of Trust Homes experiencing positive engagement behaviour are placed in cluster 2.

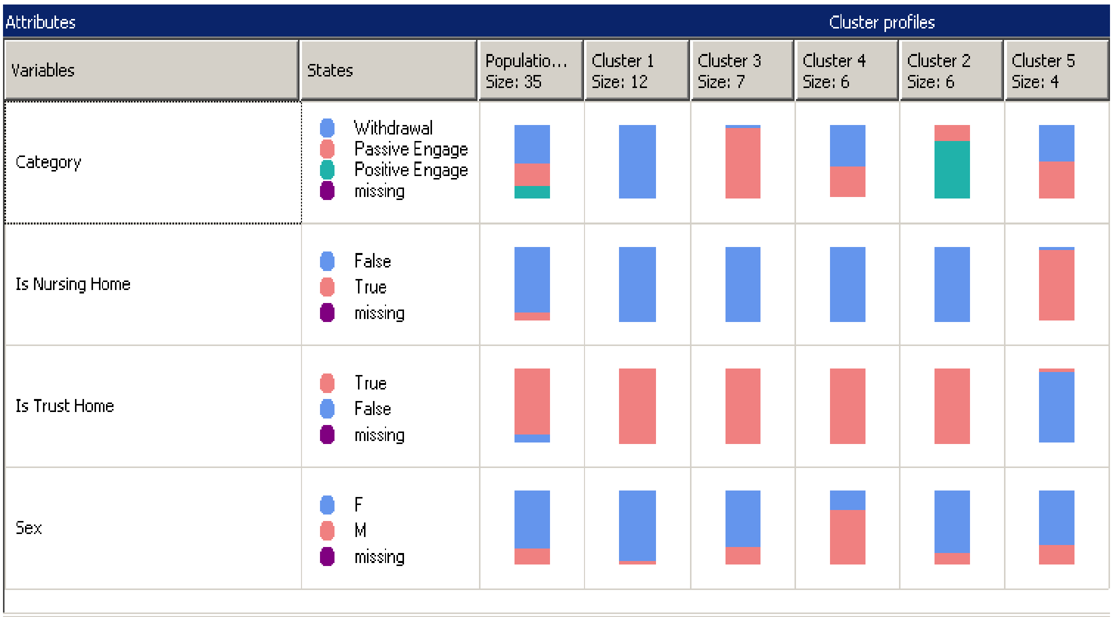


Figure 22: Cluster Profile View Showing Attributes

Figure 23 shows cluster characteristics on probability distribution of all attributes.

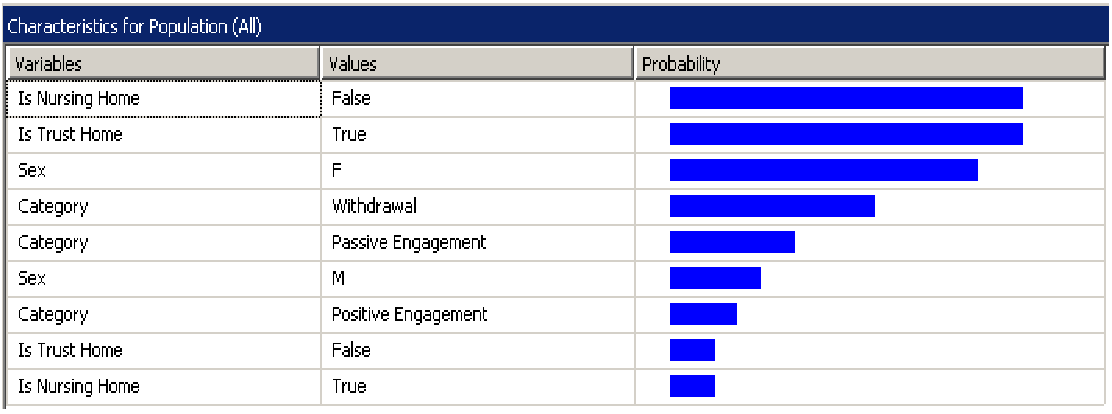


Figure 23: Cluster Characteristics of All Attributes

Probability of each attribute and its state are as shown in Table 5:

Table 5: Total Population Probability Distribution

<table>
<tr><th>Attribute</th><th>Probability Distribution</th><th>Total Probability Distribution</th></tr>
<tr><td>Is Trust Home</td><td>88.571%</td><td rowspan="2">100%</td></tr>
<tr><td>Is Nursing Home</td><td>11.429%</td></tr>
<tr><td>Category (Withdrawal)</td><td>51.429%</td><td rowspan="3">100%</td></tr>
<tr><td>Category (Passive Engagement)</td><td>31.429%</td></tr>
<tr><td>Category (Positive Engagement)</td><td>17.143%</td></tr>
<tr><td>Sex (Female)</td><td>77.143%</td><td rowspan="2">100%</td></tr>
<tr><td>Sex (Male)</td><td>22.857%</td></tr>
</table>

Overall the data set used for clustering contained the data mostly taken from Trust homes (88.571%), on female service users (77.143%) and the mostly experienced behaviour was withdrawal behaviour (51.429%). The size of the data set, number of attributes and number of attribute occurrence can affect the results of probability distribution in each cluster. The main purpose of grouping data into clusters was to explore the available amount of data to find out the classes into data for further analysis.

The EM algorithm does not compute actual assignments of objects to specific clusters but classification probabilities. In other words, each object belongs to each cluster with a certain probability. As a final result we can usually review an actual assignment of objects to clusters, based on the (largest) classification probability.

The probability distributions of each object in cluster 5 are shown in Figure 24. The service users from the Nursing Homes are showing highest probability of 95%. Female gender is prominent with highest probability of 73.8% and showing passive engagement with probability distribution of 51% and withdrawn behaviour with probability distribution of 49.2%. It can be asserted that the female service users in a Nursing Home spend most of their time either sleeping or not engaging with others.

Characteristics for Cluster 5		
Variables	Values	Probability
Is Trust Home	False	
Is Nursing Home	True	
Sex	F	
Category	Passive Engagement	
Category	Withdrawal	
Sex	M	

Figure 24: Cluster Characteristics of Cluster 5

To see what is specifically important about this cluster, we compare it with everything outside this cluster (with the other data objects from different clusters). Figure 25 shows the comparison of cluster 5 with its complements in discrimination view. It is seen that to describe the category of this cluster, the most important attribute is IsTrustHome.

Discrimination scores for Cluster 5 and Complement of Cluster 5			
Variables	Values	Favors Cluster 5	Favors Complement of Cluster 5
Is Trust Home	True		
Is Trust Home	False		
Is Nursing Home	True		
Is Nursing Home	False		

Figure 25: Cluster Discrimination of Cluster 5

Cluster 2 represents the group where all service users are from Trust Home and the majority belongs to female gender showing positive engagement during mapping sessions. (see Figure 26).

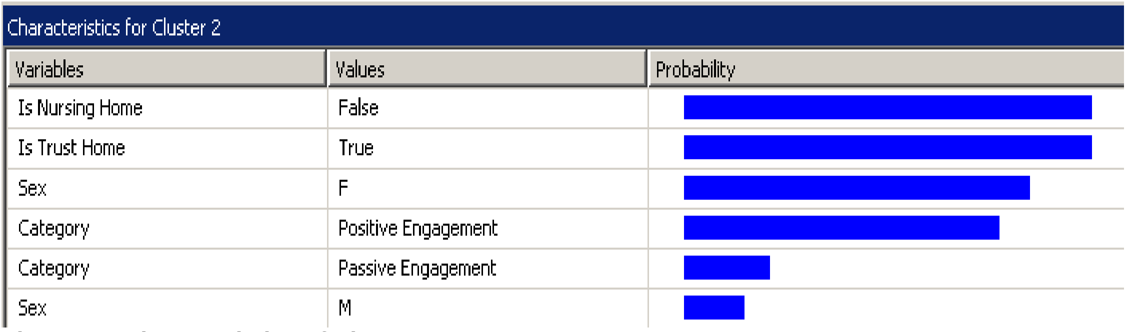


Figure 26: Characteristics of Cluster 2

6.10 Results

Clustering has been applied to a small DCM data set and it is clear that the data is grouped into different groups showing service users specific gender, behaviour and type of care setting they are in.

There was good distinction between the mean values of each of the five clusters. The interpretation of these results shows that each cluster shows the particular gender of service users and their behaviour in a particular care setting. Based on the acquired results, further data analysis can be done on individual groups to explore more hidden relationships or patterns in each group. For example Cluster 3, which has female service users with passive engagement, can be analysed further to explore the reasons for passive engagement in those females. Based on the analysis the staff of those care homes can make plans to introduce some leisure activities to change the service user's passive engagement into positive engagement.

6.11 Conclusions

I described the use of data from the DCMDW for data mining purposes. Clustering is applied to group DCM data into classes. Clustering seems to be used mostly in classification of medical data into groups for further analysis. A case study presents the grouping of the DCM data by applying clustering where data is grouped into different clusters describing the accumulation of objects in one cluster based on their similar features.

The results show that data is grouped into different clusters based on the highest probability distributions. Each cluster can be analysed further to enhance the understanding of the data. The basic purpose of this chapter was to show how the DCM data management framework (described in Chapter 5) provides the support for the use of the data from DCM data warehouse for mining purposes (for this study we used Clustering).

Depending on the availability of more DCM data other data mining tasks e.g. classification and association rules can also be applied to see hidden patterns and relationships.

Chapter 7: Conclusions and Future Work

7.1 Conclusions

To date, there has not been any published attempt nationally or internationally to design a DCM data governance framework explaining a data management solution. This interdisciplinary study describes a novel approach of managing the DCM data by applying existing IT methods and technologies.

DCM needed a governance framework based on IT technologies to manage the data nationally and internationally. The purpose of this study was to set a framework for governing DCM data to deal with data management issues from capture to dissemination. In this project we introduced the DCM data governance framework to study different steps on the data management, its quality and security. This study particularly aimed to concentrate on the data management part of the DCM data governance process where DCM data is represented at Meta level in order to allow further processing. I proposed a data warehousing approach to manage the DCM data internationally. My contributions include the designing of DCM international database and DCM data warehouse for governing the DCM data in terms of acquisition, storage and retrieval. The proposed DCM data management framework is evaluated by writing queries to extract the data from the DCM data warehouse.

Data warehouses provide a suitable environment for applying data mining techniques on the stored data as the data is pre-processed in the warehousing process. In this pre-processing step the data is checked for its quality, completeness and accuracy and arranged in structures suitable for data analysis and mining.

A case study is presented based on a limited DCM data set to describe the application of clustering on DCM data taken from the DCM data warehouse. The purpose was to identify different classes in given data and show the data warehouse and data mining connectivity. The results obtained from data warehouse queries and the case study on clustering are very encouraging and help us to see how the DCM governance framework describing the proposed data management system helps to answer the questions asked by a variety of users.

7.2 Research Contributions

This section shows the contributions made in this study, which are:

1. A design of DCM data governance framework showing the important areas of DCM to be managed. This is presented in Chapter 4. Different data governance components including DCM data management, quality and security are discussed in the context of DCM.

2. A DCM data management framework based on data warehousing approach has been proposed showing designed data management steps. The designed international DCM database and DCM data warehouse architectures are shown in Chapter 5 and published in [25].

3. A data warehousing approach to manage the DCM data facilitates the data retrieval for analysis and data mining. Different queries showing data retrieval from the data warehouse are presented in

Chapter 5. A case study is presented to show the mining application (Clustering) on DCM data in Chapter 6.

7.3 Future Work

The ideas presented in this thesis are non-exhaustive, which motivate our future work. I have been offered the opportunity to undertake further research within the School of Health Studies, University of Bradford to continue the existing MPhil work on to PhD level. During MPhil study, I found some limitations which need to be tackled in the future. The following are the recognised areas:

1) Enhancement of the DCM data warehouse structure by introducing data anonymization techniques at different security levels within the data governance framework.

2) DCM user requirements should be gathered from a variety of users to extend the DCM data warehouse architecture.

3) Enrichment of the semantics of the data warehouse schema model by creating more fact tables and dimension tables to enable designers and query optimizers to take explicit advantage of the aggregate nature of DW data.

4) The ETL process can be enhanced to meet the requirements of data update and loading issues into the DCM data warehouse.

5) This study was not meant to deal with query processing and optimization solutions, but for the future work these areimportant issues in DCM

international database and DCM data warehouse to be discussed in detail to provide a fast and efficient query response to users.

6) Data quality and security issues were out of scope of this MPhil study but are very important components of the proposed DCM data governance framework to address in detail in future.

7) Efficient methods should be applied in order to extract the data from existing DCM spreadsheets.

8) Other data mining techniques can be tested and applied on a variety of DCM data views, stored in DCM data warehouse, to find, discuss and validate hidden patterns and information from the data.

References

[1] M.F.Wisniewski, P. Kieszkowski, B.M. Zagorski, W.E. Trick, M. Sommers and R.A. Weinstein, "Development of a Clinical Data Warehouse for Hospital Infection Control ", Journal of American Medical Informatics Association, vol.10, pp. 454—462, 2003.

[2] Mental Health Foundation (2009) Dementia. [Online]. Available: http://www.mentalhealth.org.uk/information/mental-health-a-z/dementia/ [Accessed: 02 August 2010].

[3] C.P. Ferri, M. Prince, C. Brayne, H. Brodaty, L. Fratiglioni and M. Ganguli, "Global Prevalence of Dementia: A Delphi Consensus Study", The Lancet, vol. 366, no. 9503, pp. 2112-7, 2005.

[4] PSSRU "Dementia: International Comparisons". London, PSSRU Research Unit, 2007.

[5] National Dementia Strategy for England (NDS)(2010), [Online]. Available: http://alzheimers.org.uk/site/scripts/home_info.php?homepageID=111 [Accessed: 2 August 2010]

[6] Commission of the European Communities,"Communication from the Commission to the European Parliament and the Council on a European initiative on Alzheimer's Disease and other Dementias". Brussels: Commission of the European Communities, 2009.

[7] Council for the European Union, "Council Conclusions on Public Health Strategies to Combat Neurodegenerative Diseases Associated with Ageing and in Particular Alzheimer's Disease" Brussels: Council for the European Union, 2008.

[8] Norwegian Ministry of Health and Care Services, "Dementia plan 2015".Oslo: Norwegian Ministry of Health and Care Services, 2008.

[9] République Francaise, "National Plan for "Alzheimer and Related Diseases 2008-2012", Paris: République Francaise, 2008.

[10] World Health Organisation and the Milbank Memorial Fund (2000) "Towards an International Consensus on Policy for Long-Term Care of the Ageing". Geneva: World Health Organisation. [Online]. Available: http://www.milbank.org/000712oms.pdf [Accessed: 4 August 2010].

[11] Bradford Dementia Group, "Evaluating dementia care: The DCM method 7th edition". Bradford: University of Bradford, 1997.

[12] Bradford Dementia Group, "DCM 8 User's Manual". Bradford: University of Bradford 2005.

[13] D. Brooker, N. Foster, A. Banner, M. Payne and L. Jackson" The Efficacy of Dementia Care Mapping as an Audit tool: Report of a 3-year British NHS Evaluation ", Aging and Mental Health vol. 2 , no,1, pp. 60-70, 1998.

[14] G. W. Martin and D. Younger, "Anti Oppressive Practice: A Route to the Empowerment of People with Dementia through Communication and Choice", Journal of Psychiatric and Mental Health Nursing, vol. 7, pp. 59-67, 2000.

[15] D. Younger and G.W. Martin, "Dementia Care Mapping: An Approach to Quality Audit of Services for People with Dementia in Two Health Districts", Journal of Advanced Nursing, vol. 32 no. 5, pp. 1206-12, 2000.

[16] C.M. Gigliotti, S.E. Jarrott and J. Yorgason, "Harvesting Health: Effects of Three Types of Horticultural Therapy Activities for Persons with Dementia", Dementia: The International Journal of Social Research and Practice, vol. 3, no. 2, pp. 161-80, 2004.

[17] D. Kuhn, P. Edelmann and B.R. Fulton, “Daytime Sleep and Threat to Well-being of Persons with Dementia” Dementia: TheInternational Journal of Social Research and Practice, vol. 4, no. 2, pp. 233-247, 2005.

[18] D. Brooker, “Dementia Care Mapping: A Review of the Research Literature”, Gerontologist, vol. 45, (Special Issue 1), pp. 11—18, 2005.

[19] IBM Data Governance Council Maturity Model (Oct 2007), “The IBM Data Governance Council Maturity Model: Building a roadmap for effective data governance”, [Online]. Available: http://www935.ibm.com/services/uk/cio/pdf/leverage_wp_data_gov_council_maturity_model.pdf [Accessed: 15 August 2010].

[20] G. Thomas,(2009) “The Data Governance Framework”,The Data Governance Institute: Orlando, USA, [Online]. Available:http://www.datagovernance.com/dgi_framework.pdf [Accessed: 12th April 2010].

[21] J. Dyche, (June 2007), “A Data Governance Manifesto: Designing and Deploying Sustainable Data Governance”, A white paper. Baseline Consulting, [Online]. Available: http://www.siperian.com/documents/WP_JillDyche_DataGovernance_June07.pdf [Accessed: 12 April 2010].

[22] Search Data Management.com (2007), “Data Governance and Data Stewardship Strategies and Best Practice”, A book chapter, [Online]. Available: http://searchdatamanagement.techtarget.com/feature/Data-governance-and-data-stewardship-strategies-and-best-practices [Accessed: 16 April 2010].

[23] J. Salido and P. Voon, “A Guide to Data Governance for Privacy, Confidentiality, and Compliance, Part 1: A Case for Data Governance”. Microsoft Corp. USA .white paper. 2010.

[24] Online Encyclopedia (2010), [Online]. Available: http://www.encyclo.co.uk/define/scalability [Accessed: 2nd January 2010].

[25] S. Khalid, C. Surr and D. Neagu, “DCM Data Management Framework: A Data Warehousing Approach” in Information Technology in Bio-and Medical Informatics, ITBAM 2010, S.Khuri, L. Lhotska, N. Pisanti (Eds), LNCS: 6266 Proceedings, pp. 45-56.

[26] D. Brooker and C. Surr, “Dementia Care Mapping: Principles and Practice”. Bradford, University of Bradford, 2005.

[27] D.J. Power, (2003), “A Brief History of Decision Support Systems”, DSSResources.COM, World Wide Web, , version 2.8, [Online]. Available: http://DSSResources.COM/history/dsshistory.html [Accessed: 23rd March 2010].

[28] British Standard Institution(BSI), PAS 800: (2010) “Use of Dementia Care Mapping for Improved Person-centered Care in a Care Provider Organisation”.

[29] (NHS) Connecting for Health (2008), UK "Information Governance (IG)" [Online]. Available: http://www.connectingforhealth.nhs.uk/systemsandservices/infogov [Accessed: 19th April 2010].

[30] A. Bitterer and D. Newman, “Organizing for Data Quality, Gartner Research”, Stamford, CT, 2007.

[31] J.A. Osheroff, J.M. Teich, F. Blackford, B.F. Middleton, E.B. Steen, A. Wright,and D.E. Detmer, “A Roadmap for National Action on Clinical Decision Support”, Office of the National Coordinator for Health Information Technology (ONC) Report.2006.

[32] A. Lahtela and M. Hassinen, "Requirements for Radio Frequency Identification in Healthcare". Medical Informatics in a United and Healthy Europe, K.-P. Adlassnig et al. (Eds.) vol. 150, pp.720—724, 2009.

[33] D.J. Brailer, "Interoperability: the Key to the Future Health Care System". Health Affairs (Millwood) Suppl Web Exclusives, W5-19--W15-21, 2005.

[34] G.M. Hackbarth, R.D. Reichauer and M.E. Miller, "Report to the Congress: New Approaches in Medicare", Medicare Payment Advisory Commission (MEDPAC), pp. 157—181, 2004.

[35] Canada Health Infoway (2010) "Transforming Health Care", [Online]. Available: http://www2.infoway-inforoute.ca [Accessed: 06/08/2010].

[36] Canada Health Infoway (2010), "Advancing Canada's Next Generation of Health Care" [Online]. Available: http://www2.infoway inforoute.ca/Documents/Vision_2015_Advancing_Canadas_next_generation_of_health care%5B1%5D.pdf [Accessed: 07/08/2010].

[37] K. Sartipi, M.H. Yarmand and D. G. Down, "Mined-knowledge and Decision Support Services in Electronic Health" in International Conference of Software Engineering, Proceedings of the International Workshop on Systems Development in SOA Environment, 2007.

[38] NHS, Connecting for Health (2010), [Online] Available: http://www.connectingforhealth.nhs.uk/ [Accessed: 13 May 2010].

[39] R. Spronk, (2007), "The Spine an English National Programme", A white paper [Online], Available: http://www.ringholm.de/docs/00970_en.htm [Accessed: 16th May 2010].

[40] NHS, Connecting for Health, "Spine Factsheet", [Online]. Available: http://www.connectingforhealth.nhs.uk/resources/systserv/spine-factsheet [Accessed: 1st June 2010].

[41] NHS, Connecting for Health, "Secondary Uses Service (SUS)", [Online]. Available: http://www.connectingforhealth.nhs.uk/systemsandservices/sus [Accessed: 3rd June 2010].

[42] Information Society Technilogy (2010) "Biopattern network of Excellence" [Online]. Available :http://www.biopattern.org [Accessed 24rth July 2010].

[43] L. Sun, P. Hu, C. Goh, B. Hamadicharef, E. Ifeachor and I. Barbounakis et al., "Bioprofiling over Grid for eHealthcare", in First International Conference on Scalable Information Systems (INFOSCALE, 06), Hong Kong ACM Press, 2006.

[44] L. Sun, P. Hu, C. Goh, B. Hamadicharef,M.Hess and E. Ifeachor et al " Bioprofiling Over Grid for Early Detection of Dementia", ACM International Conference Proceeding Series, Proceedings of the 1st international conference on Scalable information systems, vol. 152 archive, Article 32, 2006.

[45] G.T. Henderson, E.C. Ifeachor, H.S.K. Wimalartna, E. Allen, E and N.R. Hudson, "Prospects for Routine Detection of Dementia using the Fractal Dimension of the Human Electroencephalogram", in Science, Measurement and Technology IEEE proceedings, vol. 147, no. 6, pp. 284-289, 2000.

[46] NHS, Connection for Health, "Hospital Episode Statistics (HES)", [Online]. Available: http://www.connectingforhealth.nhs.uk/systemsandservices/sus/delivery/hes/index_html [Accessed: 14th June 2010].

[47] D.J. Berndt, J.W. Fisher, J.G. Craighead, A.R. Hevner, S. Luther, S and J. Studnicki "The Role of Data Warehousing in Bioterrorism Surveillance", Decision Support Systems, vol. 43, pp.1383–1403, 2007.

[48] C.D. Ramick, "Data Warehouse in Disease Management Programmes", Journal of Health Care Information Management, vol. 15. No. 2, pp. 99—105, 2001.

[49] M.F. Wisniewski, P. Kieszkowski, B.M. Zagorski, W.E. Trick, M. Sommers, and R.A. Weinstein, "Development of a Clinical Data Warehouse for Hospital Infection Control". Journal of American Medical Information Association, vol. 10, pp. 454—462, 2003.

[50] D.J. Berndt, A.R. Hevner,and J. Studnicki, "The Catch Data Warehouse: Support for Community Health Care Decision-making", Decision Support Systems, vol. 35, pp.367–384, 2003.

[51] N. Stolba,,, "Towards a Sustainable Data Warehouse Approach for Evidence-Based Health Care", PhD Thesis (2007), Vienna University of Technology. Austria. [Online]. Available: http://publik.tuwien.ac.at/files/PubDat_141378.pdf [Accessed: 12th January 2010].

[52] M. O'Mahony and B. Van Ark (2003) "EU Productivity and Effectiveness: An Industry Perspective? Can Europe Resume the Catching Up Process?" Brussels, European Comission. [Online]. Available: http://www.fea upcam.fr/data/document_collections/38/doc_497_1188749727_fr.pdf [Accessed: 7th August 2010].

[53] E. Nolte, S. Wait and M. McKee (2006), "Investing in Health: Benchmarking Health Systems", The Nutfield Trust: For Research and Policy Studies in Health Service. [Online]. Available: http://www.nuffieldtrust.org.uk/ecomm/files/Investinginhealthweb.pdf [Accessed: 7th July 2010].

[54] NHS, Benchmarking Network (2009), "NHS Benchmarking: Raising Standards Through Sharing Excellence" [Online]. Available: http://www.nhsbenchmarking.nhs.uk/ [Accessed on 14th July 2010].

[55] NHS, "The Improvement Network: East Midlands, Benchmarking" (2010), [Online]. Available: http://www.tin.nhs.uk/local-networks/essence-of-care/background/ [Accessed: 7th July 2010].

[56] NHS, Connecting for Health (2010), "Benchmarking NHS Health Informatics Services", [Online]. Available: http://www.connectingforhealth.nhs.uk/systemsandservices/capability/phi/hottopics/benchmarking/index [Accessed: 9th July 2010].

[57] NHS, Institute for Innovation and Improvement (2007-2008), "No Better Care, Better Value Indicators", [Online]. Available: http://www.productivity.nhs.uk/IndicatorExplorer.aspx [Accessed:13 July 2010].

[58] NHS, The Information Centre for Health and Social Care, NHS Comparators (2010), [Online]. Available: http://www.ic.nhs.uk/nhscomparators [Accessed:13th July 2010].

[59] P.S. Hussey, G.F. Anderson, R. Osborn, et al, "How Does the Quality of Care Compare in Five Countries?" Health Affairs, vol. 23, pp. 89-99, 2004.

[60]. K. Pommerening (1994) "Medical Requirements for Data Protection", [Online]. Available: http://www.staff.uni-mainz.de/pommeren/Vortraege/IFIP94.pdf [Accessed: 8th March 2010].

[61] S. Schneeweiss and J. Avorn, "A Review of Uses of Health Care Utilization Databases for Epidemiologic Research on Therapeutics", A Review Article, Journal of Clinical Epidemiology, vol. 58, no. 2005, pp. 323–337, 2004.

[62] F. Rock, "Policy and Practice in the Anonymisation of Linguistic data". International Journal of Corpus Linguistics, vol. 6, no. 1, pp. 1-26, 2001.

[63] J. Heaton, "Reworking Qualitative Data", London: Sage 2004.

[64] J. Nespor, "Anonymity and Place in Qualitative Inquiry". Qualitative Inquiry, vol. 6, no. 4, pp. 546-569, 2000.

[65] J.H. Shulman, "Now You See Them, Now You Don't: Anonymity Versus Visibility in Case Studies of Teachers", Educational Researcher, vol.19, no. 6, pp. 11-15, 1990.

[66] D. Thomson, L. Bzdel, K. Golden-Biddle, T. Reay and C.A. Estabrooks, "Central Questions of Anonymization: A Case Study of Secondary Use of Qualitative Data", Forum: Qualitative Social Research, vol. 6, no 1, Article no 29 &# 8211, 2005.

[67] Canadian Institutes of Health Research (CIHR) (April 2003). "Executive Summary of Secondary Use of Personal Information in Health Research: Cases Studies", Ottawa, Canada: CIHR, [Online]. Available: http://www.cihr-irsc.gc.ca/e/about/6827.shtml, [Accessed: 21st May 2010].

[68] L. Corti, A. Day and G. Backhouse, (2000), "Confidentiality and Informed Consent: Issues for Consideration in the Preservation of and Provision of Access to Qualitative Data Archives [46 paragraphs]", Forum Qualitative Sozialforschung / Forum: Qualitative Social Research [Online], vol.1 (3), Art. 7. Available: http://www.qualitative-research.net/fqs-texte/3-00/3-00cortietal-e.htm. [Accessible: 3rd January 2010]

[69] Shared Pathology Informatics Network (SPIN) (2004), National Cancer Institute (NCI), USA [Online]. Available: http://spin.nci.nih.gov/ [Accessed: [21st July 2010]

[70] J.J. Berman, J. J, "Confidentiality Issues for Medical Data Miners", Artificial Intelligence in Medicine, Elsevier Science, vol. 26 (2002), pp. 25–36, 2000.

[71] P. Samarati and L. Sweeney, L, "Generalizing Data to Provide Anonymity when Disclosing Information", in Proceedings of the 17th ACM SIGMOD-SIGACT-SIGART, Symposium on the Principles of Database System, vol.188, 1998.

[72] R. Bayardo and R. Agrawal, "Data Privacy Through Optimal k-anonymization" in Proc. of the 21st Int'l Conference on Data Engineering, IEEE Computer Society Washington DC, pp. 217-228, 2005.

[73] G. Aggarwal,T. Feder, K. Knthapadi, R. Motwani, R. Panigrahy, D. Thomas, D and An Zhu, "Approximation Algorithms for k-Anonymity", Journal of Privacy Technology, in Proceedings of the 10th International Conference on Database Theory (ICDT'05), 2005.

[74] A. Meyerson and R. Williams, R, "On the Complexity of Optimal k-anonymity" In Proceeding.of the 23rd ACM SIGMOD- SIGACT-SIGART, Symposium on the Principles of Database Systems, pp. 223-228, 2004.

[75] J. Byun, A. Kamra, E. Bertino and N. Li" Efficient k-Anonymization Using Clustering Techniques", R. Kotagiri et al. (Eds.): DASFAA 2007, LNCS 4443, pp. 188–200, Springer-Verlag Berlin Heidelberg, 2007.

[76] A. Campan, T.M. Truta, J. Miller and R. Sinca, "A Clustering Approach for Achieving Data Privacy", 2007.

[77] K. LeFevre, D.J. DeWitt and R. Ramakrishnan, "Incognito: Efficient FullDomain KAnonymity", ACM-SIGMOD 1416, Baltimore, Maryland, USA, 2005.

[78] J. Liu and K. Wang, "Anonymizing Transaction Data by Integrating Suppression and Generalization", LNCS, Advance in Knowledge Discovery and Data Mining, vol. 6118 (2010), pp. 171-180, Springer Berlin / Heidelberg, 2010.

[79] B.A. Malin, "Protecting DNA Sequence Anonymity with Generalization Lattices", Methods of Information in Medicine, vol. 44, no. 5, pp. 687-692, 2005.

[80] G. Szarvas, R. Farkas and R. Busa-Fekete, “State-of-the-art Anonymization of Medical Records Using an Iterative Machine Learning Framework”, Journal of American Medical Information Association, vol. 14, no. 5, pp. 574–580, 2007.

[81] The HIPAA Guide- Security and Privacy Rules (2007), [Online]. Available: http://www.hipaaguide.net/ [Accessed: 10th August 2010].

[82] C. Newcomb, (2008), “Implementing Data Governance in Complex Healthcare Organisations: Challenges and Strategies”, A white paper, Baseline Consulting, [Online]. Available: http://baselineconsulting.web5.hubspot.com/Default.aspx?app=LeadgenDownload&shortpath=downloads%2fbcg-wp-governance-healthcare.pdf [Accessed: 5th June 2010].

[83] M. R. Chassin and R.W. Galvin,“The Urgent Need to Improve Health Care Quality”. Institute of Medicine National Roundtable on Health Care Quality. Journal of American Medical Association, vol. 280, no.11, 1998.

[84] D. E. Detmer, “Information Technology for Quality Health Care: A Summary of United Kingdom and United States experiences”, Special Article, Qual Health Care, vol. 9, no. 3, pp. 181-189, 2000.

.

[85] B. Chaudhry, J. Wang, S. Wu, M. Maglione, W. Mojica, E. Roth, S.C. Morton and P. G. Shekelle (2006) “Systematic Review: Impact of Health Information Technology on Quality, Efficiency, and Costs of Medical Care Annals of Internal Medicine”, Improving Patient Care, vol. 144, no 10, pp. 742-752, [Online]. Available: http://www.annals.org/content/144/10/742.full.pdf+html [Accessed: 15th July 2010].

[86] Canada Health Act Annual Report (2007- 2008), [Online]. Available: http://www.healthcanada.gc.ca/medicare [Accessed: 12th February 2010].

[87] W. Raghupathi and J. Tan, “Strategic IT Applications in Health Care”, Communications of the ACM, vol. 45, no. 12, pp. 56-61, 2002.

[88] R. Jacoby, “The Emerging Role of Electronic Medical Records-What Do They Mean to Quality and Value?”, vol. 2, no. 2, 2008, [Online]. Available: http://jdc.jefferson.edu/cgi/viewcontent.cgi?article=1035&context=vbp [Accessed: 13th July 2010]

[89] K. Weber, B. Otto and H. Osterle, “One Size Does not Fit All- A Contingency Approach to Data Governance”, ACM Journal of Data and Information Quality, vol. 1, no. 1, Article. 4, (2009).

[90] NHS, Connecting for Health (2008) "Information Governance (IG)", [Online]. Available: http://www.connectingforhealth.nhs.uk/systemsandservices/infogov [Accessed: 19th April 2010].

[91] A. Berson and L. Dubov, (2007), “Master Data Management and Customer Data Integration”, Data Management Concerns of MDM-CDI Architecture, Chapter 6. pp. 107-132, [Online]. Available: http://media.techtarget.com/searchDataManagement/downloads/MDM_and_CDI_ch6.pdf [Accessed: 12th April 2010].

[92] T. Kennedy, “Governance, Risk, and Compliance”, Sapinsider/ Wellesly Information Science (WIS), 2007.

[93] K. Wende, “A Model for Data Governance – Organising Accountabilities for Data Quality Management”, Proceedings of Australasian Conference on Information Systems. Toowoomba, 2007.

[94] W.H. Inmon, “Building the Data Warehouse”, 4th ed. John Wiley and Sons, 1996.

[95] D. Ross, “Retail Data Warehousing The- State-of-the-Art, In Global Coverage of the Business Intelligence Ecosystem”, US ed. (2005), [Online]. Available: http://www.b-eye-network.com/view/769[Accessed: 11th January 2010].

[96] The Data Warehouse Institute Awards: ARC’s Fraud Prevention Solutions, TDWI Best Practices Awards, (2009), [Online]. Available: http://www.arccorp.com/news/pr20090623.pdf [Accessed: 2nd January 2010].

[97] N. Stolba, A. M. Tjoa, M. Thomas and B. Marko, “Federated Data Warehouse Approach to Support the National and International Interoperability of HealthCare Information Systems”, 15th European Conference on Information systems (ECIS), St Gallen, Switzerland (2007)

[98] D. J. Protti, “How Business Intelligence is Making Health Care Smarter”, World View Report (10) (2005), [Online]. Available: http://www.connectingforhealth.nhs.uk/newsroom/worldview/protti10 [Accessed: 24th February 2010].

[99] S.Khalid, “A Web-Based Data Management System for an International Data Repository on the Quality of Dementia Care”, MSc Thesis 2009, University of Bradford, UK,

[100] T. Connolly and C. Begg, “Database Systems, A Practical Approach to Design, Implementation, and Management”. 4th ed, England: Pearson Education Limited,1999.

[101] J. Germany, J. Walker, T. Clark, “Foundation Of Database Design Series”, Logical Database Design Principles, 1st ed. Auerbach Publications, 2005.

[102] S. Chauhuri and U, Dayal, “An Overview of Data Warehousing and OLAP Technology”, ACM SIGMOD archive, vol.26, no. 1, pp. 65-74, 1997.

[103] P. Gray and H. Watson, Decision Support in the Data Warehouse, Prentice-Hall, Englewood Cliffs, NJ, 1998.

[104] P, Vassiliadis, A, Simitsis and S, Skiadopoulos, “Conceptual Modeling for ETL Processes”, in Proceedings of the 5th ACM international workshop on Data Warehousing and OLAP (DOLAP, 02), McLean, Virginia, USA, 2002.

[105] J. Trujillo and M. Luján-Mora, A UML Based Approach for Modeling ETL Processes in Data Warehouses, In Conceptual Modeling – ER 2003, LNCS, Vol. 2813, pp. 307--320. Springer Berlin / Heidelberg, 2003.

[106] P. Ponniah, “Data Warehousing Fundamentals: A Comprehensive Guide for IT Professionals” John Wiley & Sons, Inc, 2001.

[107] S. Anderson,(2010), “What is Data Scrubbing?”, [Online]. Available: http://www.wisegeek.com/what-is-data-scrubbing.htm [Accessed: 15th August 2010].

[108] Mathurin Body, Maryvonne Miquel, Yvan Bédard and Anne Tchounikine, “A multidimensional and multiversion structure for OLAP applications”, in Proceedings of the 5th ACM international workshop on Data Warehousing and OLAP, DOLAP, 02, McLean, Virginia, USA pp. 1-6, 2002.

[109] A. Gupta, V. Harinarayan and D. Quass, “Aggregate-Query Processing in Data Warehousing Environment”, Proceedings of 21st VLDB Conference, Zurich: Switzerland, 1995.

[110] A. Shoshani, “OLAP and Statistical Databases: Similarities and Differences”, ACM Transaction on Database Systems, SIGART symposium on Principles of database, 1997.

[111] An ESRI White Paper (1998) “Spatial Data Warehousing for Hospital Organizations”, Environmental Systems Research Institute, Inc. USA. [Online]. Available: http://www.esri.com/library/whitepapers/pdfs/sdwho.pdf [Accessed: 26th July 2010].

[112] J.L. San and K. Siau, “A Review of Data Mining Techniques”, Industrial Management and Data Systems, MCB University Press, vol. 101, no.1, pp. 41-46, 2001.

[113] M. Silver, T. Sakata, H.C. Su, C. Herman, S.B. Dolins and M. J. O’Shea, “Case Study: How to Apply Data Mining Techniques in a Healthcare Data Warehouse”, Journal of Healthcare Information Management, vol. 15, no. 2, 2001.

[114] U. Fayyad, G. Piatetsky-Shapiro and P. Smyth, P, “From Data Mining to Knowledge Discovery: An Overview’’, Fayyad, U., Piatestsky-Shapiro, G., Smyth, P. and Uthurusamy, R. (Eds), Advances in Knowledge Discovery and Data Mining, MIT Press, Cambridge, MA. 1996.

[115] M. Ester, H. Kriegel, J. Sander, M. Wimmer and X. Xu, X. “Incremental Clustering for Mining in a Data Warehousing Environment” in Proceedings of the 24th VLDB Conference, New York: USA, 1998.

[116] R. Agrawal, T. Imielikski and A. Swami, A, “Mining Association Rules between Sets of Items in Large Databases” Proceeding of the ACM SIGMOD Conference. Washington DC: USA, 1993.

[117] A. K. Jain, M.N. Murty and P.J. Flynn, P.J, “Data Clustering: A Review. ACM Computing Surveys, vol. 31, no 3, 1999.

[118] J. Han and M. Kamber, M. “Data Mining: Concepts and Techniques. 2nd ed. Elsevier: Morgan Kaufmann, 2006.

[119] C. Fraley and A.E. Raftery, “How Many Clusters? Which Clustering Method?Answers via Model-Based Cluster Analysis”. Technical Report no 329, Seattle: USA

[120] J. Kolsky (1997), “Interpreting Cluster Analysis”, [Online]. Available: http://www.camo.com/rt/Resources/infodoc/Interpreting_Cluster_Analysis.html [Accessed: 12th August 2010].

[121] StatSoft, Inc. (2010). Electronic Statistics Textbook. Tulsa, OK: StatSoft. [Online]. Available: http://www.statsoft.com/textbook [Accessed: 17th August 2010].

[122] S. Hirano and S. Tsumoto, "Cluster Analysis of Time-Series Medical Data Based on the Trajectory Representation and Multiscale Comparison Techniques", icdm, Sixth IEEE International Conference on Data Mining (ICDM'06), pp. 896-901, 2006.

[123] D. Dilts, J. Khamalah, A. Plotkin, “Using Cluster Analysis for Medical Resource Decision Making”, Medical Decision Making: Sage Journal Online, vol. 15, no. 4, pp. 333-347, 1995.

[124] D. Greene, A. Tsymbal, N. Bolshakova and P. Cunningham, “Ensemble Clustering in Medical Diagnostics”, 17th IEEE Symposium on Computer-Based Medical Systems (CBMS'04), 2004.

[125] M. Easter, H. P. Kriegel, J. Sander and X. Xu, “ A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise” in Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining (KDD 96), 1996.

[126] OPTICS, “Ordering Points To Identify the Clustering Structure”, [Online]. Available: http://en.wikipedia.org/wiki/OPTICS_algorithm [Accessed: 1st July 2010].

[127] J. Rogers, (2001), “Data Mining Using the EM Clustering Algorithm on Places Rated Almanac Data”, [Online]. Available:

http://www.galaxy.gmu.edu/stats/syllabi/inft979/RogersPaper.pdf [Accessed: 8th August 2010].

Appendix A: Glossary

DCM
Dementia Care Mapping

Service User
A person with dementia who is residing in a care home or attends the Day Care

Participant
A service user who is being observed (mapped) by a Mapper for a specific time period

Unit
A care home or a hospital ward where mapping takes place

Mapper
A person who maps the participants over a specific time period and provides the feedback reports to the unit staff.

Researcher
A person who can be a student, a practitioner or anyone who wants to carry the research study on DCM system

BCC
Behavior Category Code

ME
Mood and Engagement Code

PD
Personal Detractions

PE
Personal Enhancers

Time Frame
A specific time period (usually 5min) to write the behavior and mood of the participant

Primary Care Trust (PCT)

PCT is a type of NHS trust, part of the National Health Service in England, that provides some primary and community services or commission them from other providers, and are involved in commissioning secondary care.

DW
Data Warehouse

DCMDW
Dementia Care Mapping Data Warehouse

OLAP
Online Analytical Application Processing

OLTP
Online Transactional Processing

DM
Data Mining

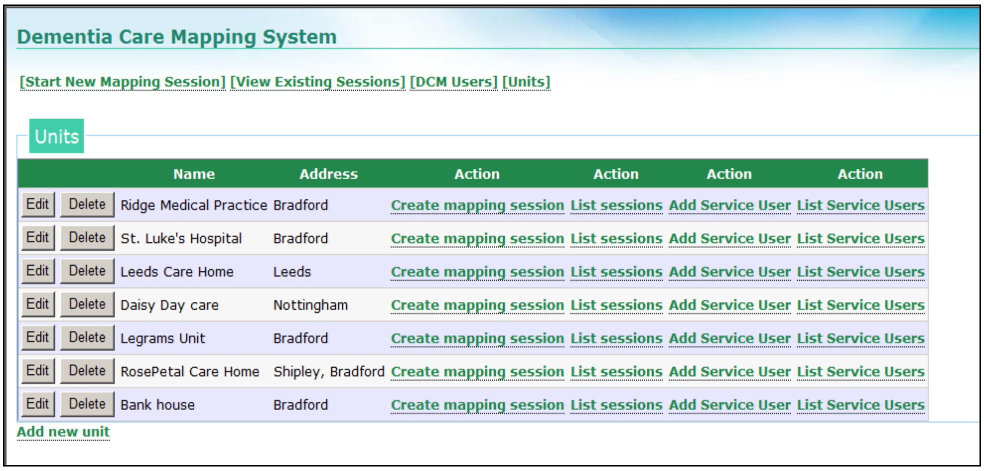
Appendix B: A list of Mood and Engagement (ME) values

Mood	ME value	Engagement
Very happy, cheerful. Very high positive mood.	+5	Very absorbed, deeply engrossed/engaged
Content, happy, relaxed. Considerable positive mood	+3	Concentrating but distractible. Considerable engagement
Neutral. Absence of overt signs of positive or negative mood	+1	Alert and focused on surroundings. Brief or intermittent engagement
Small signs of negative mood	-1	Withdrawn and out of contact
Considerable signs of negative moods	-3	
Very distressed. Very great signs of negative mood.	-5	

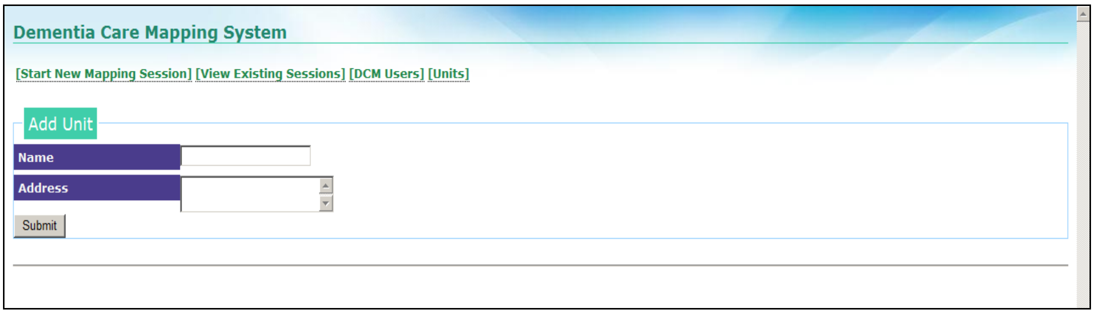
Appendix C: A list of Behaviour Category Codes

Code	Memory Cue	General Description of Category
A	Articulation	Interacting with others verbally or otherwise-with no obvious accompanying activity.
B	Borderline	Being engaged but passively
C	Cool	Being disengaged, withdrawn
D	Doing for self	Self care
E	Expressive	Expressive or creative activity
F	Food	Eating or drinking
G	Going back	Reminiscence and life review
I	Intellectual	Prioritising the use of intellectual activities.
J	Joints	Exercise or physical sport
K	Kum and go	Walking, standing or moving independently
L	Leisure	Leisure, fun and recreational activities
N	Nod Land of	Sleeping, dozing
O	Objects	Displaying attachment to or relating to inanimate objects
P	Physical	Receiving practical, physical or personal care
R	Religious	Engaging in a religious activity
S	Sexual expression	Sexual expression
T	Timalation	Direct Engagement of the sense
U	Unresponded to	Attempting to communicate without receiving a response
V	Vocational	Work or work like activity
W	Withstanding	Repetitive self-stimulation of a sustained nature
X	X-cretion	Episode related to excretion
Y	Yourself	Interaction in the absence of any observable other
Z	Zero option	Fits none of the existing category

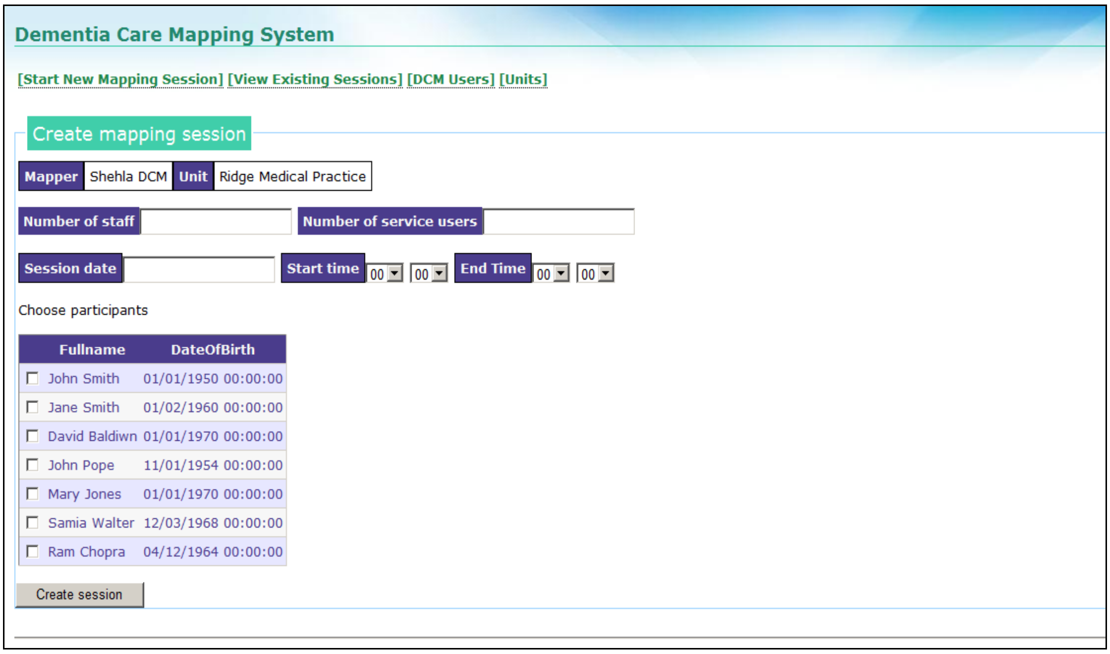
Appendix D: Screen Shots of Web-based Interface Developed for DCM International Database



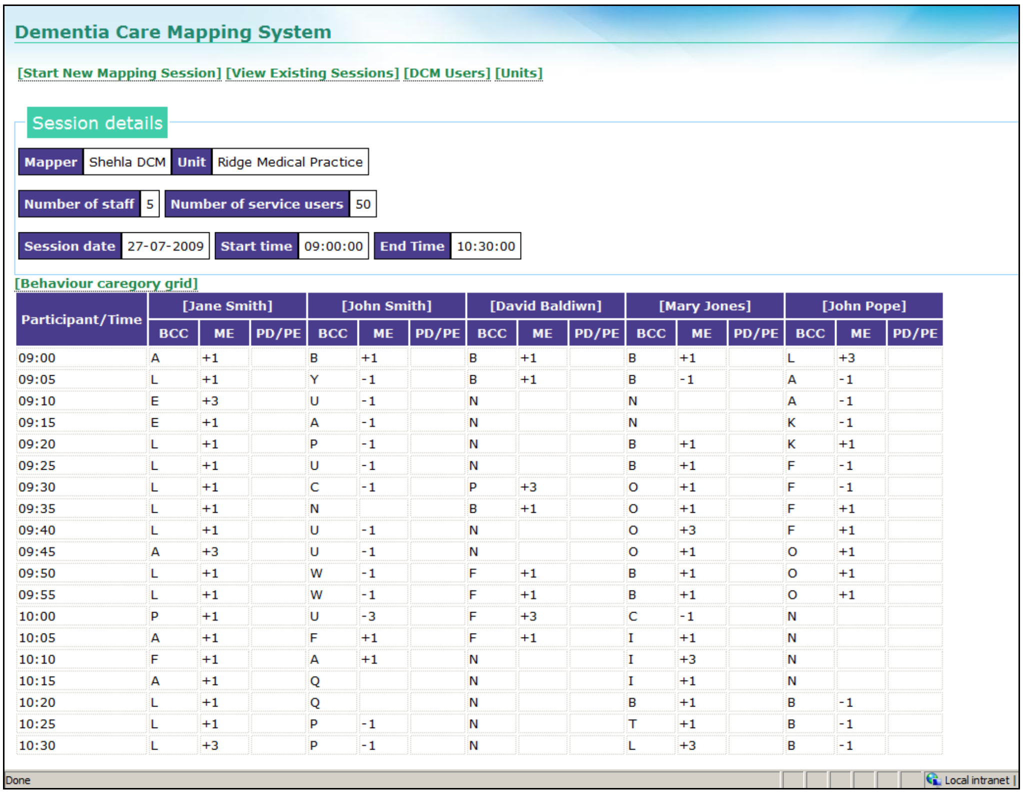
Dementia Care Mapping (DCM) Main Page



DCM Add Unit Page

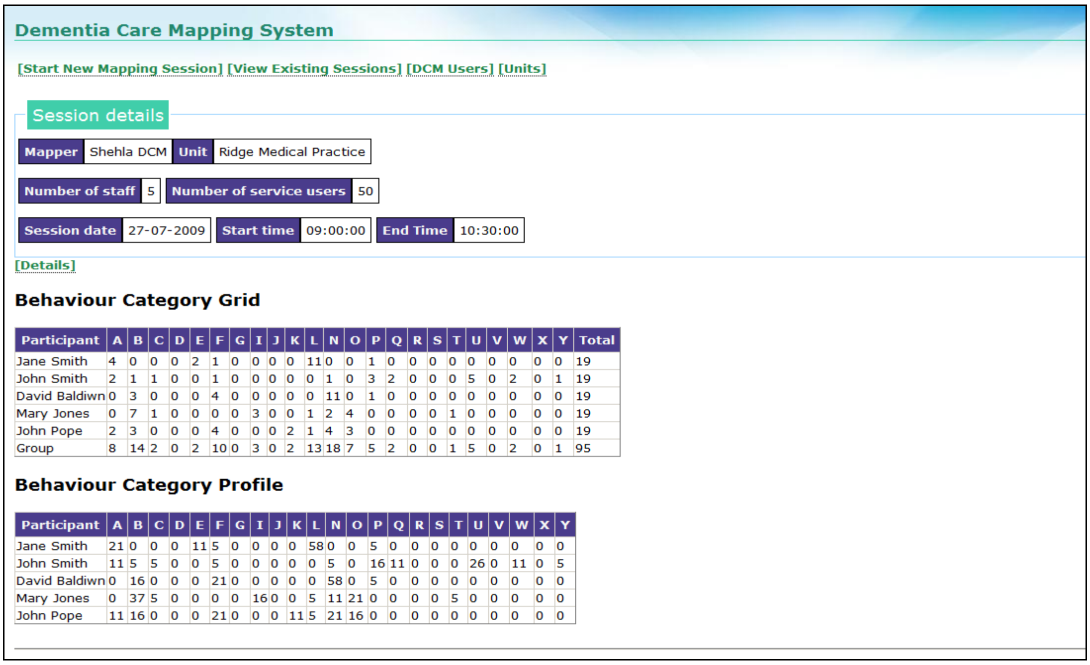


Create Mapping Session Page



Participant/Time	[Jane Smith]			[John Smith]			[David Baldiwn]			[Mary Jones]			[John Pope]		
	BCC	ME	PD/PE	BCC	ME	PD/PE	BCC	ME	PD/PE	BCC	ME	PD/PE	BCC	ME	PD/PE
09:00	A	+1		B	+1		B	+1		B	+1		L	+3	
09:05	L	+1		Y	-1		B	+1		B	-1		A	-1	
09:10	E	+3		U	-1		N			N			A	-1	
09:15	E	+1		A	-1		N			N			K	-1	
09:20	L	+1		P	-1		N			B	+1		K	+1	
09:25	L	+1		U	-1		N			B	+1		F	-1	
09:30	L	+1		C	-1		P	+3		O	+1		F	-1	
09:35	L	+1		N			B	+1		O	+1		F	+1	
09:40	L	+1		U	-1		N			O	+3		F	+1	
09:45	A	+3		U	-1		N			O	+1		O	+1	
09:50	L	+1		W	-1		F	+1		B	+1		O	+1	
09:55	L	+1		W	-1		F	+1		B	+1		O	+1	
10:00	P	+1		U	-3		F	+3		C	-1		N		
10:05	A	+1		F	+1		F	+1		I	+1		N		
10:10	F	+1		A	+1		N			I	+3		N		
10:15	A	+1		Q			N			I	+1		N		
10:20	L	+1		Q			N			B	+1		B	-1	
10:25	L	+1		P	-1		N			T	+1		B	-1	
10:30	L	+3		P	-1		N			L	+3		B	-1	

View DCM Data Page



Dementia Care Mapping System

[Start New Mapping Session] [View Existing Sessions] [DCM Users] [Units]

Session details

Mapper	Shehla DCM	Unit	Ridge Medical Practice

Number of staff	5	Number of service users	50

Session date	27-07-2009	Start time	09:00:00	End Time	10:30:00

[Details]

Behaviour Category Grid

Participant	A	B	C	D	E	F	G	I	J	K	L	N	O	P	Q	R	S	T	U	V	W	X	Y	Total
Jane Smith	4	0	0	0	2	1	0	0	0	0	11	0	0	1	0	0	0	0	0	0	0	0	0	19
John Smith	2	1	1	0	0	1	0	0	0	0	0	1	0	3	2	0	0	0	5	0	2	0	1	19
David Baldiwn	0	3	0	0	0	4	0	0	0	0	0	11	0	1	0	0	0	0	0	0	0	0	0	19
Mary Jones	0	7	1	0	0	0	0	3	0	0	1	2	4	0	0	0	0	1	0	0	0	0	0	19
John Pope	2	3	0	0	0	4	0	0	0	2	1	4	3	0	0	0	0	0	0	0	0	0	0	19
Group	8	14	2	0	2	10	0	3	0	2	13	18	7	5	2	0	0	1	5	0	2	0	1	95

Behaviour Category Profile

Participant	A	B	C	D	E	F	G	I	J	K	L	N	O	P	Q	R	S	T	U	V	W	X	Y
Jane Smith	21	0	0	0	11	5	0	0	0	0	58	0	0	5	0	0	0	0	0	0	0	0	0
John Smith	11	5	5	0	0	5	0	0	0	0	0	5	0	16	11	0	0	0	26	0	11	0	5
David Baldiwn	0	16	0	0	0	21	0	0	0	0	0	58	0	5	0	0	0	0	0	0	0	0	0
Mary Jones	0	37	5	0	0	0	0	16	0	0	5	11	21	0	0	0	0	5	0	0	0	0	0
John Pope	11	16	0	0	0	21	0	0	0	11	5	21	16	0	0	0	0	0	0	0	0	0	0

Behaviour Category Code Page